

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
12 December 2002 (12.12.2002)

PCT

(10) International Publication Number
WO 02/099725 A1

- (51) International Patent Classification⁷: **G06F 19/00**
- (74) Agent: **MYERS BIGEL SIBLEY & SAJOVEC**; PO Box 37428, Raleigh, NC 27627 (US).
- (21) International Application Number: PCT/US02/16406
- (22) International Filing Date: 23 May 2002 (23.05.2002)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/296,018 5 June 2001 (05.06.2001) US
60/356,616 13 February 2002 (13.02.2002) US
10/145,521 13 May 2002 (13.05.2002) US
- (71) Applicant (*for all designated States except US*): **IN-CELLICO, INC.** [US/US]; Suite 205, 2327 Englert Drive, Durham, NC 27713 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (*for US only*): **WILBANKS, John**, Thompson [US/US]; 8101 Reynard Road, Chapel Hill, NC 27516 (US). **LEVY, Joshua**, Lerner [US/US]; 4523 Oak Hill Road, Chapel Hill, NC 27514 (US). **SEGARAN, Suresh, Toby** [NZ/US]; 700 Bolinwood, #16F, Chapel Hill, NC 27514 (US). **GARDNER, Richard, N.** [US/US]; 10101 Daviton Court, Raleigh, NC 27615 (US).
- (81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.



WO 02/099725 A1

(54) Title: **SYSTEMS, METHODS AND COMPUTER PROGRAM PRODUCTS FOR INTEGRATING BIOLOGICAL/CHEMICAL DATABASES TO CREATE AN ONTOLOGY NETWORK**

(57) Abstract: Biological/chemical databases are integrated by obtaining an entity-relationship model for each of the biological/chemical databases, and identifying related entities in the entity-relationship models of at least two of the biological/chemical databases. At least two of the related entities that are identified are linked, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases. The entity-relationship model that integrates the biological/chemical databases provides an ontology network that integrates the diverse ontologies that are represented by the independent biological/chemical databases. By navigating the entity-relationship model in response to queries, discovery may be obtained that may not be obtainable from any one of the independent biological/chemical databases.

**SYSTEMS, METHODS AND COMPUTER PROGRAM PRODUCTS FOR
INTEGRATING BIOLOGICAL/CHEMICAL DATABASES TO CREATE AN
ONTOLOGY NETWORK**

Cross Reference to Provisional Applications

This application is related to and claims the benefit of U.S. Application Serial No. _____ to Wilbanks, Levy, Segaran and Gardner, filed May 13, 2002, entitled *Systems, Methods and Computer Program Products for Integrating*
5 *Biological/Chemical Databases to Create an Ontology Network* (Attorney Docket 9223-10), which itself is related to and claims the benefit of Provisional Application Serial No. 60/296,018 to Levy and Segaran, filed June 5, 2001, entitled *Cell: A Cross-Referenced Ontological Database for Biological Data*; and Provisional
10 Application Serial No. 60/356,616 to Gardner and Wilbanks, filed February 13, 2002, entitled *Ontology Networks, a New Foundation for Discovery*, all of which are assigned to the assignee of the present application, the disclosures of all of which are hereby incorporated herein by reference in their entirety as if set forth fully herein.

Field of the Invention

15 This invention relates to bioinformatics/cheminformatics, and more particularly to systems, methods and computer program products for processing biological databases and/or chemical databases.

Background of the Invention

20 The biotechnology, chemical and pharmaceutical industries continue to attempt to develop innovative and effective drugs, chemicals, agricultural and/or other products on shorter schedules and at reduced cost. A potential challenge faced in this pursuit is managing the enormous volume, diversity and complexity of data that is currently being generated by these industries. In particular, new technologies have
25 resulted in an enormous increase in the amount of data available to researchers. Unfortunately, this enormous increase in the amount of data may not lead to

corresponding advances in discovery, because the sheer volume of data may outpace the ability of researchers to transform that data into knowledge.

In an attempt to analyze these massive amounts of data, the field of bioinformatics has emerged. See, for example, U.S. Patent Application Serial No. 09/657,218 to Wilbanks et al., filed September 7, 2000, entitled *Systems, Methods and Computer Program Products for Processing Genomic Data In An Object Oriented Environment*, assigned to the assignee of the present application, the disclosure of which is hereby incorporated herein by reference in its entirety as if set forth fully herein.

The massive volume of data that is being generated also may be accompanied by a large diversity of data sources that may generate the data. For example, public, private, proprietary, clinical, chemical, genomic and other databases from various data sources may be produced. Unfortunately, it may be difficult to integrate these heterogeneous data sources.

One conventional approach for data integration uses a data warehouse and data mining techniques. A data warehouse may use a relational database and a star model in which searchable database fields are stored in their own tables, forming a star around a table of records. Unfortunately, it may be difficult to integrate new types of data without significant modification to the table structure. Moreover, querying the assembled information using conventional data mining techniques also may present potential problems. These queries may range in sophistication from simple use of Boolean operators, data search engines such as Internet-based search tools, and/or more sophisticated query languages that employ relational inquiries into the database. Unfortunately, these queries may require significant knowledge of the data sources, the structure of the assembled data, and/or experience in the use of query languages. The use of Internet-based search engines may yield inaccurate yet exhaustive reams of information that may not be relevant to the original request.

Another conventional approach that may be used for data integration is the flat-file or link-driven federation, wherein users can perform text searching on the databases independently, and then jump to different databases, for example via World Wide Web links. Although a flat-file or link-driven federation may simplify searching for non-expert users, it may be difficult to search across multiple databases simultaneously. Moreover, it may be difficult to obtain desired information for data records that only are indirectly and/or inferentially linked.

Another conventional integration technique is referred to as a wrapper or view, which can provide cross-database querying without moving data from the original databases. For each database, a separate driver may be designed that can query the database. A wrapper can then ask several databases for some results and bring them
5 together to find intersections. Unfortunately, it may be difficult to bring in new data types, as new drivers may need to be provided for every new data source. Moreover, queries may be slow and memory-intensive, because all relevant databases may need to be queried for their entire result set before elimination by any other parts of the query is performed. Finally, relationships may not be provided unless specified in the
10 queries and/or wrappers.

Summary of the Invention

Some embodiments of the present invention integrate a plurality of biological/chemical databases by obtaining an entity-relationship model for each of
15 the plurality of biological/chemical databases, and identifying related entities, including identical entities, in the entity-relationship models of at least two of the biological/chemical databases. At least two of the related entities that are identified are linked, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases. In some embodiments, when the entities are
20 identical entities, they are merged. In some embodiments, each of the plurality of databases represents an ontology and the entity-relationship model that integrates the plurality of biological/chemical databases creates an ontology network.

Accordingly, ontology networks according to some embodiments of the present invention can link related entities in entity-relationship models of independent
25 biological/chemical databases, to thereby create a single entity-relationship model for the independent biological/chemical databases. By navigating the single entity-relationship model in response to queries, discovery may be obtained that may not be obtainable from any one of the independent biological/chemical databases.

In some embodiments, linking is performed by merging at least two of the
30 identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases. In other embodiments, merging is accomplished by establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical

databases, a respective alias of which refers to a respective one of the identical entities that are identified.

In some embodiments, the traversing is performed from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending
5 entity. In other embodiments, the entities are traversed from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity. In yet other embodiments, the entities are traversed in response to a query and in response to at least one path rule. In some embodiments, the at least one path rule specifies the type of path to use in traversing through the plurality of entities, the type
10 of path not to use in traversing through the plurality of entities, the type of ending entity that can be included in the query results, the type of ending entity that is not to be included in the query results, the type of relationship to be used in traversing through the plurality of entities, the type of relationship that is not to be used in traversing through the plurality of entities and/or a confidence level to be achieved in
15 traversing through the plurality of entities. In still other embodiments, groups of relationships may be classified into a class of relationships, and the at least one path rule can specify a class of relationships to be included or excluded. Multiple classes can be assigned to a given relationship.

In other embodiments, the query results are stored as at least one new
20 relationship in the entity-relationship model that integrates the plurality of biological/chemical databases, to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of biological/chemical databases. In still other embodiments, a confidence level is assigned to at least one of the relationships in the entity-relationship model that
25 integrates the plurality of biological/chemical databases. In still other embodiments, query results also may be based on assigned confidence levels.

According to other embodiments of the present invention, a new biological/chemical database may be integrated with a plurality of biological/chemical databases, by providing an entity-relationship model of the plurality of
30 biological/chemical database that links at least some related entities in at least two of the biological/chemical databases. An entity-relationship model for the new biological/chemical database is obtained. Related entities in the entity-relationship model of the new biological/chemical database and the entity-relationship model of the plurality of biological/chemical databases are identified. At least two of the

related entities that are identified are linked, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database. In other embodiments, the entity-relationship model of the plurality of biological/chemical databases that links at least some related entities
5 in the at least two of the biological/chemical databases provides an ontology network and the entity-relationship model of the new biological/chemical database represents an ontology.

In other embodiments of the invention, when linking identical entities, the at least two of the identical entities that are identified are merged into a single entity in
10 the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database. In other embodiments, merging may be accomplished by establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database. A respective alias refers to a respective
15 one of the at least two of the identical entities that are identified.

In other embodiments, the new biological/chemical database is an updated version of one of the plurality of biological/chemical databases. In some of these embodiments, at least one entity is identified that is in the one of the plurality of biological/chemical databases and that has been deleted from the updated version of
20 the one of the plurality of biological/chemical databases. An alias that is associated with the at least one entity is removed. In still other embodiments, at least one entity is split based upon the alias that was removed. In yet other embodiments, an image of the at least one record that has been deleted may be retained in the plurality of biological/chemical databases, so as to allow an archival history to be maintained. In
25 still other embodiments, multiple images or instances of the entity/relationship structure may be maintained to reflect updates and/or deleted records and/or query results, and these multiple instances may be correlated to one another to obtain new knowledge.

In still other embodiments, when adding a new biological/chemical database,
30 entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database are identified. At least one new entity is added to the entity-relationship model that

corresponds to the entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model.

Bioinformatics data processing systems according to some embodiments of the present invention include an ontology network engine that is configured to build
5 an integrated entity-relationship model of a plurality of independent biological/chemical databases. The entity-relationship model comprises a plurality of entities including links and also comprises a plurality of relationships. In some embodiments, a metadata database is configured to store therein the integrated entity-relationship model of the plurality of independent biological/chemical databases. In
10 other embodiments, a loader is configured to load an independent entity-relationship model of each of the independent biological/chemical databases into the ontology network engine. The independent biological/chemical databases may be loaded in a typeless format. Other embodiments include a virtual experiment layer that is configured to conduct virtual experiments on the integrated entity-relationship model.
15 Yet other embodiments include a discovery layer that is configured to discover biological/chemical knowledge from the integrated entity-relationship model. Moreover, in still other embodiments, the integrated entity-relationship model provides a bioinformatics data structure. Finally, it will be understood that any of the embodiments described herein may be provided as systems, methods and/or computer
20 program products.

Brief Description of the Drawings

Figures 1 and 2 illustrate conceptual overviews of environments in which some embodiments of the present invention may be used.

25 Figure 3 is a hardware/software block diagram of some embodiments of the present invention.

Figure 4 is a software architecture diagram of some embodiments of the present invention.

Figure 5 is a flowchart of operations for integrating biological/chemical
30 databases according to some embodiments of the present invention.

Figure 6 is a flowchart of operations for integrating a new biological/chemical database into a plurality of biological/chemical databases according to some embodiments of the present invention.

Figure 7 is a flowchart of operations for querying a plurality of biological/chemical databases according to some embodiments of the present invention.

Figure 8 is an example of a portion of an entity-relationship data structure that integrates multiple biological/chemical databases according to some embodiments of the present invention.

Figure 9 is a flowchart of operations for integrating biological/chemical databases according to some embodiments of the present invention.

Figure 10 is a flowchart of operations for integrating new biological/chemical databases according to some embodiments of the present invention.

Figure 11 is a flowchart of operations for performing queries according to some embodiments of the present invention.

Figures 12-17 conceptually illustrate an example of the creation of an ontology network according to some embodiments of the present invention.

Figure 18 illustrates an example of querying an ontology network that was created in Figures 12-17 according to some embodiments of the present invention.

Figure 19 illustrates another example of an ontology network that may be created according to some embodiments of the present invention.

Figure 20 is an example of linkages that may be provided by an ontology network of Figure 19 according to some embodiments of the present invention.

Figure 21 illustrates a browser display of a portion of an ontology network according to some embodiments of the present invention.

Figure 22 is a block diagram of a data processing architecture that may be used with some embodiments of the present invention.

Figures 23A and 23B, which together form Figure 23, is an entity-relationship diagram of a conceptual schema for an ontology network according to some embodiments of the present invention.

Figures 24 and 25 are flowcharts of operations for integrating biological/chemical databases and integrating new biological/chemical databases according to some embodiments of the present invention.

Figure 26 is a flowchart illustrating operations for traversing an ontology network using path rules according to some embodiments of the present invention.

Figure 27 is an example of an *in silico* experiment that can be derived from an ontology network according to some embodiments of the present invention.

Figures 28-35 illustrate an example of a path rule that may be used to obtain discovery according to some embodiments of the present invention.

Figure 36 illustrates an example of a display screen that may be used to initiate a query using a path rule that was specified in Figures 28-35 according to some
5 embodiments of the present invention.

Figures 37A and 37B, which together form Figure 37, illustrates an example of a display screen of query results that may be obtained according to some embodiments of the present invention.

Figures 38 and 39 are flowcharts of operations for querying an ontology
10 network according to some embodiments of the present invention.

Figure 40 illustrates a conceptual overview of environments in which some embodiments of the present invention may be used.

Figures 41 and 42 illustrate examples of ontology networks that can be used to link personal data, securities data and government data according to some
15 embodiments of the present invention.

Detailed Description of Preferred Embodiments

The present invention now will be described more fully hereinafter with reference to the accompanying figures, in which embodiments of the invention are
20 shown. This invention may, however, be embodied in many alternate forms and should not be construed as limited to the embodiments set forth herein.

Accordingly, while the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however,
25 that there is no intent to limit the invention to the particular forms disclosed, but on the contrary, the invention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the invention as defined by the claims. Like numbers refer to like elements throughout the description of the figures.

The present invention is described below with reference to block diagrams and/or flowchart illustrations of methods, apparatus (systems) and/or computer
30 program products according to embodiments of the invention. It is understood that each block of the block diagrams and/or flowchart illustrations, and combinations of blocks in the block diagrams and/or flowchart illustrations, can be implemented by computer program instructions. These computer program instructions may be

provided to a processor of a general purpose computer, special purpose computer, and/or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer and/or other programmable data processing apparatus, create means for implementing the functions/acts specified in the block diagrams and/or flowchart block or blocks.

These computer program instructions may also be stored in a computer-readable memory that can direct a computer or other programmable data processing apparatus to function in a particular manner, such that the instructions stored in the computer-readable memory produce an article of manufacture including instructions which implement the function/act specified in the block diagrams and/or flowchart block or blocks.

The computer program instructions may also be loaded onto a computer or other programmable data processing apparatus to cause a series of operational steps to be performed on the computer or other programmable apparatus to produce a computer-implemented process such that the instructions which execute on the computer or other programmable apparatus provide steps for implementing the functions/acts specified in the block diagrams and/or flowchart block or blocks.

It should also be noted that in some alternate implementations, the functions/acts noted in the blocks may occur out of the order noted in the flowcharts. For example, two blocks shown in succession may in fact be executed substantially concurrently or the blocks may sometimes be executed in the reverse order, depending upon the functionality/acts involved.

Definitions

As used herein, the following terms have the following meanings:

Biological/chemical: Biological and/or chemical.

Biological database: A database that includes, at least in part, data describing or related to biological experiments and/or concepts at any number of biological levels, from population to organism to gene and/or protein sequence. Examples include, but are not limited to, the well known KEGG, MaizeDB, OmIm and HGMD databases. Biological databases can include genomic databases that include, at least in part, data containing genome sequence and/or data related to genome sequence such as annotation and/or gene prediction. Examples of genomic databases include, but are not limited to, the well known ENSEMBL, WormPep and Celera Human

Genome databases. Biological databases can also include proteomic databases that include, at least in part, data from or related to proteomic experiments, such as 2d-gel, or results from high-throughput mass spectrometry. Examples of proteomic databases include, but are not limited to, the well known Swiss-2D-PAGE database. Biological
5 databases also can include classification databases, examples of which include, but are not limited to, the well known MeSH, Gene Ontology Consortium (GO) and Enzyme databases. Biological databases also can include sequence databases that include, at least in part, data containing biomolecule sequence information, such as nucleotide, peptide and/or carbohydrate sequence and/or annotation. Examples of
10 sequence databases include, but are not limited to, the well known GenBank and SWISS-PROT databases. Biological databases also can include toxicity, disease, clinical trial and/or other databases that describe or relate to biological experiments and/or concepts at any number of biological levels, from population to organism to gene and/or protein sequence.

15 Chemical database: A database that includes, at least in part, chemical information such as chemical structures, formulae, nomenclature, properties and/or biochemical action of organic and/or inorganic chemicals. Examples include, but are not limited to, the well known ChemID+ database.

Entity-relationship: A data model that views information as a set of basic
20 objects (entities) and relationships among these entities. An entity is an object or concept about which information is stored. An entity may have attributes which are the properties or characteristics of the entity. Relationships indicate how two entities share information. Relationships may also have attributes or properties. The entity-relationship model was originally developed by Dr. Peter P. Chen and was adopted as
25 the meta model for the American National Standards Institute (ANSI) Standard on Information Resource Directory System (IRDS). In a biological/chemical database, examples of entities include, but are not limited to, 2D-gel-spot, carbohydrate, chemical, classification, disease, express-in, gene, gene-product, interaction, keyword, literature, localization, locus, motif, nucleotide-sequence, oligonucleotide, pathway,
30 physical location, protein, symbol, taxonomy, related-group, marker, pseudogene, strain, tissue-cell-type, variation, reaction, clone, experiment, experiment-result, structure and sequence-library, and examples of relationships include kind of, reaction type, default, path, reaction-left, reaction-right, annotation, available oligonucleotide, catalysis, enzyme classification, enzyme in pathway, expression, glycosylation,

homology, homology cluster, in species, inheritable locus, isomer, kind-of, mapping, marker, nomenclature, occurs-in, ontological, part of complex, part of experiment, part of interaction, part of pathway, part of structure, partial-sequence, protease-class, protein contains motif, pseudogene, reaction type, reference, related, result probe, same gene product, same protein, sequenced, spot contains, transcription, translation, variation, in strain, reactant, library sequence, exon-gene-annotation, exon-sequence-annotation, and mapped-between.

Ontology: A structured vocabulary of terms and some specification of their meaning and/or relationships among one another based on a set of beliefs about the terms and their meanings/relationships. The structure can be explicit and/or implicit.

Other terms used herein have their ordinary meaning to those having skill in the art, unless specified otherwise, and, therefore, need not be expressly defined herein.

Referring now to Figure 1, a conceptual overview of environments in which embodiments of the present invention may be used, is shown. As shown in Figure 1, these environments may include large amounts of data from many biological/chemical experiments 102 that may be collected in many disparate or independent databases including public, private, proprietary, clinical, chemical, genomic and/or other databases 104. Each database may have associated therewith a quality control tool 106 that can check for errors, database integrity and/or other parameters within the individual database.

Still referring to Figure 1, data mining tools may be used as were described above, to allow searching within and/or across databases 104. However, data mining/data warehousing may have shortcomings in integrating and/or querying diverse databases. Moreover, in other embodiments, data mining tools need not be used.

Still referring to Figure 1, some embodiments of the present invention may provide knowledge mining, using aliases and/or ontology networks, wherein a plurality of biological/chemical databases is integrated, so that new knowledge may be established by querying the integrated data structure. This knowledge mining can lead to the running of virtual experiments 112, also referred to as *in silico* experiments, using the integrated databases and one or more virtual experiment tools. These virtual experiments 112 then can lead to new discoveries 114 which may be

obtained using one or more discovery tools. Accordingly, embodiments of the present invention can provide a knowledge mining layer 110 that can allow virtual experiments 112 and discovery 114, respectively, to be obtained, based on independent biological/chemical databases 104 that are collected from disparate sources.

Referring now to Figure 2, another conceptual overview of environments in which embodiments of the present invention may be used is shown. As shown in Figure 2, a plurality of disparate biological/chemical databases may be provided. For example, a genomic/proteomic database 202a, a biomolecule database 202b, a phenotypic database 202c, a public database 202d and a curated/third party database 202n may be provided. More or fewer databases also may be provided, and one or more of these databases may be merged or bifurcated.

Each of these databases 202a-202n includes records for a plurality of biological/chemical objects, also referred to herein as entities. These databases 202a-202n also generally include an indication of one or more relationships among the various biological/chemical objects, to thereby define an entity-relationship data structure or model for each of the independent databases. The entity-relationship data structure for each database may be thought of as defining an ontology, which provides a vocabulary of terms and some specification of their meaning and/or relationships among one another. These entities and relationships may represent a set of beliefs on the part of the database creator or other individual(s)/organization(s). Thus, the ontology in a given database 202a-202n represents a belief system about the entities and relationships of the data in the database. Some of the databases 202a-202n may constitute a relational database data model that does not explicitly contain entity-relationship data structures. However, entity-relationship data models may be derived from these data models using conventional techniques, in some embodiments of the invention. In other relational database models, one or more entities may be present or derivable, but relationships may not be present or implicit in the data models. According to some embodiments of the invention, these data models can be integrated with other databases that include an ontology, to provide an ontological context for the data model as well.

Referring again to Figure 2, the databases 202a-202n may constitute a data collection layer that may be derived from, for example, wet laboratory experiments. Some of this data may be processed in a quality control layer by data analysis/quality

control modules 204a, 204b...204n. These data analysis/quality control modules may provide some data curation and determination of clusters of meaningful information. Other databases, such as databases 202d and 202n, may not include an analysis/quality control layer.

5 Still referring to Figure 2, in some embodiments, at least some of the raw, compressed and/or qualified data may be incorporated into a warehouse by a data integration/data mining layer 206, which can enable the organization of the data into logically structured tables of information. Data querying may conventionally be performed at the data integration/data mining tool or layer 206, for example by
10 developing specialized query requests to gain inference or knowledge from the warehouse. In other embodiments, a data integration/data mining tool 206 is not used.

 In some environments, embodiments of the present invention may operate on top of this data integration/data mining tool 206, and/or may also operate directly on a biological/chemical database, such as the chemical data and cheminformatics database
15 208, and/or the pre-clinical database 214. The preclinical database 214 may include ADME, toxicity, pharmaco-kinetics and/or other data. Some embodiments of the present invention can provide a knowledge mining layer in the form of an ontology network 210 that can overlay/merge/associate diverse ontologies that are represented in diverse databases, data tables and/or data repositories. The resulting ontology
20 network 210 thus can link multiple disparate ontologies.

 As will be described in more detail below, according to some embodiments of the present invention, an ontology network 210 can incorporate the entity-relationship models of the databases on which it is built, but can also define new relationships or hierarchies by the process of overlay, merge and/or association of entities from the
25 independent ontologies. This conceptualization of knowledge can serve as a specification mechanism for the development of a broad-mesh belief system that can deliver experimental insight. Stated differently, ontology networks 210 according to some embodiments of the present invention can traverse and, thereby, establish a linked path of relationships creating associations between characteristically unlike
30 entities, to thereby allow the revelation of new information and knowledge. The resulting lattice of semantically rich metadata can form an ontology network 210 that captures the knowledge from the data sources 202, 208 it supports.

 Thus, as shown in Figure 2, in some embodiments of the present invention, an ontology network 210 can be located above the data integration layer 206, and can

provide a knowledge tool or layer that is available for hypothesis or question-driven mining, as opposed to complex data mining queries that may be typical of data mining applications. Thus, some embodiments of the invention can provide a meta-database of entities and/or relationships that can allow efficient and intelligent analysis of
5 accumulated data.

Still referring to Figure 2, ontology networks 210 according to some embodiments of the present invention may be linked to an application tool or layer, such as a discovery/prediction and simulation tool 212, so as to allow more accurate discovery, prediction and/or simulation. Examples of a discovery/prediction and
10 simulation layer 212 are described in Provisional Application Serial No. 60/346,694 to Segaran and Pan, filed January 7, 2002, entitled *Analysis of Functional Cellular Pathways and the Role of Structural Homology on Chemosensitivity*, the disclosure of which is hereby incorporated herein by reference in its entirety as if set forth fully herein.

15 Referring now to Figure 3, a hardware/software block diagram of some embodiments of the present invention now will be described. It will be understood that some embodiments of the present invention may execute on one or more personal, application and/or enterprise computer systems, in a standalone, networked, distributed, pervasive, peer-to-peer and/or other configuration.

20 Referring now to Figure 3, a data processing engine 300, which also may be referred to as an ontology engine, can be used to integrate, update and/or query a plurality of databases, and/or generate, add to and/or query an ontology network as will be described in detail below. The engine 300 can provide a knowledge mining layer 110 of Figure 1 and/or an ontology network 210 of Figure 2 in some
25 embodiments. The engine 300 is responsive to one or more loaders 302 that can extract relevant information from one or more biological/chemical databases 304, which can be analogous to the data collection layer 104 of Figure 1 and/or the databases 202, 208 of Figure 2. In some embodiments, *a priori* knowledge of the semantics of the ontology that is represented by the associated biological/chemical
30 databases 304 is built into the loader 302 of that ontology's external data files. Moreover, in some embodiments, the loader 302 has knowledge of the semantics of the appropriate part of the engine 300, to which the ontology data connects.

In some embodiments, the engine 300 generates metadata in the form of an overlaid/merged/associated entity-relationship data structure, which can be stored in a

metadata database 308. One or more applications 306 may be used for providing discovery, prediction, simulation and/or other applications, analogous to the discovery layer 114 of Figure 1 or the discovery/prediction and simulation layer 212 of Figure 2. These applications 306 can interface with a local user interface and/or can interface with a Web browser 316 that is connected to a Web server 312, for example, via a network, such as the Internet 314. The design of a Web server 312, a network such as the Internet 314, and a Web browser 316 is well known to those having skill in the art and need not be described further herein. Finally, user-defined path rules 322 and/or predefined path rules 324 may be provided to allow directed path traversals as will be described in detail below.

Figure 4 is a software architecture diagram of some embodiments of the present invention. These embodiments may be used on one or more personal, application and/or enterprise computer systems in a standalone, networked, distributed, pervasive, peer-to-peer and/or other configuration. As shown in Figure 4, a data processing engine 400 can generate the metadata for a metadata database 408 as will be described in detail below. An Application Programming Interface (API) 430 may be provided to interface the engine 400 with one or more external database loaders 402 and one or more applications 406. The engine 400, metadata database 408, loaders 402 and applications 406 may be analogous to elements 300, 308, 302 and 306, respectively, of Figure 3.

Referring now to Figure 5, operations for integrating biological/chemical databases according to some embodiments of the present invention now will be described. It will be understood that these operations may be embodied, for example, in a knowledge mining layer 110 of Figure 1, an ontology network 210 of Figure 2, an engine 300 of Figure 3 and/or an engine 400 of Figure 4. These embodiments can integrate a plurality of disparate or independent biological/chemical databases, such as the databases 202a-202n and 208 of Figure 2, and/or 304 of Figure 3, each of which includes records for a plurality of biological/chemical objects.

Referring now to Block 502, a set of records is identified in the plurality of biological/chemical databases that relates to (i.e., is associated with) a single biological/chemical object. At Block 504, an entity is established in a data structure that corresponds to the single biological/chemical object. The entity includes a plurality of aliases, a respective one of which refers to a respective record in the set of records in the plurality of biological/chemical databases. At Block 506, if there are

more records, the operations for identifying and establishing (Blocks 502 and 504, respectively), are repeatedly performed for a plurality of sets of records and, in some embodiments, for all sets of records, in the plurality of biological/chemical databases, to establish a plurality of entities in the data structure.

5 Still referring to Figure 5, in other embodiments of the invention, as shown at Block 510, the plurality of entities in the data structure are linked in an entity-relationship model of the plurality of biological/chemical databases. It will be understood that the operations of Block 510 may be performed in parallel with the operations of Block 504, and need not be performed after a plurality or all sets of
10 records have been identified (Block 502) and entities have been established (Block 504).

 Still referring to Figure 5, according to other embodiments of the invention, at Block 512, a query may be received. The query may be received from an application or other program with or without direct user intervention. As shown at Block 514, the
15 query may identify or specify a path type through the entity-relationship model. As shown at Block 516, in some embodiments, if no path type is identified, the plurality of entities that are linked in an entity-relationship model is traversed in response to a query, to thereby obtain query results that are based on the records in the plurality of biological/chemical databases. In contrast, at Block 518, if a path type is identified,
20 the plurality of entities that are linked in an entity-relationship model is traversed along the identified type of path or paths in response to a query, to thereby obtain query results that are based on the records in the plurality of biological/chemical databases. These query results may be provided at Block 520 via an application, such as an application tool 306 of Figure 3 and/or 406 of Figure 4. These queries may
25 provide virtual experiments and/or discovery (Blocks 112 and 114 of Figure 1), and/or discovery/prediction and simulation (Block 212 of Figure 2). These queries also may represent discovery processes that are recorded and reused.

 As will be described in detail below, in some embodiments, the query may specify a starting entity and an ending entity, and the operations of Block 516 can
30 traverse the plurality of entities that are linked in the entity-relationship model from the starting entity to the ending entity, to thereby identify relationships between the starting entity and the ending entity that are based on the entity-relationship model of the plurality of biological/chemical databases. In other embodiments, the entities are traversed from a starting entity to a plurality of ending entities in response to a query

that specifies the starting entity, to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the entity-relationship model of the plurality of biological/chemical databases.

Moreover, the path type of Block 514 may be identified using one or more path rules, such as user-defined path rules 322 and/or predefined path rules 324 of Figure 3. The path rules may specify, for example, a type of path to use in traversing through the plurality of entities, a type of path not to use in traversing through the plurality of entities, a type of ending entity that can be included in the query results, a type of ending entity that is not to be included in the query results, a type of relationship to be used in traversing through the plurality of entities, a type of relationship that is not to be used in traversing through the plurality of entities and/or a confidence level to be achieved in traversing through the plurality of entities. Many other path rules also may be provided.

Finally, when the query results are provided in the Block 520, some embodiments store the query results that are based on the entity-relationship model of the plurality of biological/chemical database, as at least one new relationship is the entity-relationship model. Knowledge that was derived from the query thereby may be stored in the entity-relationship model.

Referring now to Figure 6, operations for integrating a new biological/chemical database into a plurality of biological/chemical databases, each of which includes records for a plurality of biological/chemical objects, according to some embodiments of the present invention, now will be described. At Block 602, a data structure is provided that includes a plurality of entities, a respective one of which corresponds to a single biological/chemical object. At least some of the entities include a plurality of aliases, a respective one of which refers to a record in a respective one of the plurality of biological/chemical databases that relates to a single biological/chemical object. In some embodiments, the operations of Block 602 may be provided by performing the operations of Blocks 502-510 in Figure 5. Thus, a preexisting data structure may be provided, and/or a data structure may be generated as was described in Figure 5.

Referring again to Figure 6, at Block 604, records are identified in the new biological/chemical database that correspond to at least one of the entities in the existing data structure. In some embodiments, the new biological/chemical database includes an entity-relationship model or an entity-relationship model is generated

therefor. In other embodiments, the new database may merely be a relational database data model that does not, explicitly or implicitly, define relationships. By integrating the entity or entities in this new database with the existing entity-relationship model, an ontological context can be provided for the new database.

5 Then, at Block 606, aliases are added to at least one of the entities of the data structure that correspond to the records in the new biological/chemical database, to thereby integrate the new biological/chemical database into the plurality of biological/chemical databases. Thus, additional biological/chemical databases may be readily integrated into the data structure for a plurality of biological/chemical
10 databases.

Referring again to Figure 6, in other embodiments of the invention, operations may be provided for identifying when a record in the new biological/chemical database corresponds to two or more entities in the existing data structure (Block 608). If this is the case, then at Block 610, the two or more entities in the existing
15 data structure are merged into a new entity that includes aliases that correspond to the records associated with the two or more entities in the data structure, as well as the record in the new biological/chemical database that corresponds to the two or more entities in the data structure. Thus, the data structure can be modified as new databases are incorporated.

20 Still referring to Figure 6, operations may be performed according to other embodiments of the present invention, when the new biological/chemical database is an updated version of one of the plurality of biological/chemical databases that already are contained in the data structure. Thus, as shown at Block 612, at least one record in the one of the plurality of biological/chemical databases that has been
25 deleted from the updated version of the one of the plurality of biological/chemical databases is identified. At Block 614, when such a record has been identified, the at least one record is removed from the one of the plurality of biological/chemical databases that has been deleted. At Block 616, aliases that are associated with the at least one record also are removed. Moreover, at Block 618, the at least one entity in
30 the data structure may be split based upon the aliases that were removed. Thus, as new versions of one or more of the databases are incorporated to replace an older version, the data structure may be updated.

In yet other embodiments of the invention, when the data structure is updated by addition, deletion and/or splitting, an image, instance or version of the earlier data

structure may be maintained. This image may be used for archival purposes, to ascertain the state of the data structure during a discovery, according to some embodiments of the invention. In other embodiments, comparisons may be made between different images of the data structure, to itself lead to new discovery. Thus, for example, one image of the entity-relationship model can store data related to successful drug discoveries, from genomic to clinical indicators, to extract traversal patterns related to likelihood of success. Another image can store a similar set of patterns for expensive drug failures that did not make it through a genomic, pre-clinical or clinical phase. These images can be compared in order to obtain discovery that can predict success.

Referring now to Figure 7, operations for querying a plurality of biological/chemical databases, each of which includes records for a plurality of biological/chemical objects, now will be described according to some embodiments of the present invention. As shown in Figure 7 at Block 602, a data structure including a plurality of entities and a plurality of aliases, is provided, as already was described in connection with Figure 6. Then, the plurality of entities that are linked in an entity-relationship model is traversed in response to a query, to thereby obtain query results, for example using operations 512-520 of Figure 5. These operations will not be described again for the sake of brevity.

Additional qualitative discussion of integration and/or querying of biological/chemical databases according to some embodiments of the present invention that were described in Figures 5-7 now will be provided. In particular, some embodiments of the invention can import different types of experimental, sequence, chemical, annotation, or other data from a Tab-Separated-Value (TSV) format, a simple eXtensible Markup Language (XML) format and/or other formats. Scripts may be provided to convert all common data formats to this TSV, XML and/or other formats. Some embodiments can create biological entities with many different aliases, parents and children. Entities can be merged if they are found to be equivalent. The entities may be organized in Directed Weighted Graph (DWG) based ontologies, as well as hierarchical and/or single level classifications. For non-expert users, a HyperText Markup Language (HTML)-based database viewer, which allows the user to search for terms and then move between different entities via hyperlinks, may be provided. Other embodiments also can produce a tool for traversing across multiple relationships to construct a logical path. Yet other embodiments can provide

a tool for importing stored traversals in order to automatically execute those traversals across multiple entities.

Thus, some embodiments of the invention can provide a cross-reference query tool for searching across multiple databases, returning only entities which meet the specified query criteria in all databases. Other embodiments also can provide a translation and annotation tool that can allow translation from one naming system to another naming system, and automatic annotation of data files using different naming systems with description data from differing imported databases. Still other embodiments can provide a clustering engine and viewer, which can allow a user to take clustered experimental data from another program and compare it with data clustered by differing data types (e.g., molecular function) to see how well the experimental clusters predict the annotation clusters and if there are additional annotation clusters. Finally, still other embodiments can provide an unsupervised grouping search, which can take a list of clustered biological entities (e.g., genes showing a similar expression pattern) and can automatically generate a hypothesis of why they are grouped.

Accordingly, some embodiments of the present invention can bridge the naming system barrier by acquiring information from databases with names of entities residing in multiple repositories, and merging one or many entities as appropriate. Heretofore, lack of merging may have been a barrier to query expansion. In particular, biological research often includes the understanding that a natural and intuitive relationship exists between components of biological entities, such as a cell, cell walls, genes, proteins, sequences, etc., and these relationships can be documented to provide a mechanism to build a traversal across multiple such entities, to establish an interpreted or inferred solution. These traversals also can identify a cause and effect relationship. Embodiments of the invention can merge the different names of the identical entities from different unintegrated (independent) data repositories, to thereby allow these traversals to be accomplished. Thus, embodiments of the present invention can apply an integration layer above the disparate data repositories and, therefore, can bind many related data repositories together. These embodiments can enable and promote increased biological context and information mining.

Some embodiments of the invention can generate, expand, update and/or query a data structure containing many nodes, each representing a biological entity (such as a protein, a gene, a protein family, or a literature reference) with multiple

aliases. Using biological entity nodes, rather than a different table for each database (as in a star schema), means that all records in diverse biological/chemical databases that represent the same object can be merged into a single entity. For example, many "integrated" databases, include a table of SWISS-PROT records and a table of PIR records, which would be joined by a reference point or hub. A cross-reference in the SWISS-PROT entry may indicate that it is the same protein as a PIR entry. In contrast, in some embodiments of the invention, these records are used to create a single biological entity, label it with a category "protein" and establish aliases from both SWISS-PROT and PIR so it can be referenced using either naming system.

10 In other embodiments, the entities or nodes are connected by relationships into a DWG, which means that every entity can have multiple children and multiple parents. Because there are so many categorization methods for biological entities such as genes and proteins, there may be a need for multiple non-identical groupings for an entity. The DWG allows a single entity to be grouped with other entities by as many different methods as desired, while still allowing these groups to be kept
15 separate from each other.

In other embodiments, the data structure is also designed to be typeless, meaning that, although each entity is associated with a specific category, the same data structure can be used to represent all entities, as well as relationships between them. By using the same data structure, the data structure can potentially store any
20 type of data without any modification. Moreover, some embodiments of the present invention can traverse the DWG unsupervised, so that these embodiments do not need to be told which path to take in order to find relationships or similarities.

Some embodiments of the invention may be implemented in both object oriented and Relational Database Management Systems (RDBMS) models, each of
25 which may have potential advantages. One of the potential advantages of a relational database is that it may be queried with Structured Query Language (SQL). Also, since potential users may already own an RDBMS, deployment can be simpler. If a user does not own an RDBMS there are many systems available. A potential
30 advantage of an object oriented database implementation is that interaction with object-oriented software can be simpler than with an RDBMS.

Figure 8 is an example of a portion of an entity-relationship data structure that can integrate multiple biological/chemical databases according to some embodiments of the invention. In Figure 8, the entities or nodes, represented by the ovals, contain a

quoted string specifying their category (e.g., "gene"). The lines between the nodes indicate parental relationships (also referred to as group membership), with the parent groups displayed higher in Figure 8. The text items connected to the entities are their aliases, which show the naming system (e.g. EMBL, SWISS-AC) and the identifier within that naming system. There are two proteins in Figure 8, and both are referenced by the same Medline article. However, only the protein on the right of Figure 8 has an associated Pfam domain. Below the proteins in Figure 8 are the genes that translate to the protein.

As was described above, some embodiments of the present invention can identify and merge records in a plurality of biological/chemical databases that represent the same entity. Since identifiers within a naming system are considered to be unique, two objects with the same naming system-identifier pair are considered to be identical. In some embodiments, as was described in connection with Blocks 608 and 610, a record will be added and have an identity cross-reference, also referred to as an alias, to a record that has already been incorporated. When an alias is attached to an entity, some embodiments of the invention can check if the exact naming system-identifier pair is already in use. If it is, the entities are merged together, creating a new entity with all of the relationships, aliases and properties of its component entities.

It also will be understood that databases that are integrated according to some embodiments of the invention can be updated often, in some cases weekly or even daily. If new records are added to the databases, embodiments of the invention can add more entities, aliases and/or relationships. Other embodiments may remove or delete references or entries from databases as was described in Blocks 612-618. Deletion may not be explicit - that is to say, there may be nothing in the data file that states, "Entry ABC was removed". Instead, the entry may not be present in a subsequent version of the database. Some database vendors, (e.g., GCG's SeqStore product) may approach this issue by rebuilding the entire database with the new data on a regular basis. Unfortunately, this can break relationship links to private annotations that the user might have added, and may even remove these annotations altogether. The total rebuild also may be time-consuming.

According to some embodiments of the invention, deletion may be handled by tagging every alias and every relationship with the database from which it came (the source) and the date of its last update. When a record is read in, some embodiments

of the invention can find the entity to which it points and can check the aliases and relationships to see if any of them have the same source as this record. If any aliases or relationships are found which have the same source, but are not in this record, it is determined that they were removed from the record (Block 612) and they can be
5 removed from the database (Blocks 614 and 616) without the need to impact the data that came from other sources.

Moreover, according to other embodiments of the invention, when deleting a record/alias, a situation may occur where two entities had been merged because of a cross-reference, but this cross-reference is later deleted. In this case, some
10 embodiments of the invention may need to determine whether or not to split the entity into several other entities, and which aliases each should have (Block 618). This determination can be thought of as a graph theory problem, which can be solved by determining the transitive closure of the aliases (as nodes) and the update information (as connections). The existence of a connection between two aliases can be used as
15 an indication that they belong in the same entity. If all the aliases belong in the same entity then a split may not need to be made.

The following Examples shall be regarded as merely illustrative and shall not be construed as limiting the invention. These Examples represent data management problems for which some embodiments of the invention may be used. In the
20 Examples, a description is provided of how one may approach the problem using embodiments of the invention, a link-federated database and a data warehouse. In these Examples, the user may be a bench scientist with a vague understanding of bioinformatics, but with no programming or database administration skills.

25 Example 1 - Translation

The user is experimenting with bonobo apes. There is a bonobo ape database (BonoboBase), which is not in the user's database, but the user has a table of links (BonoboToGenpept.txt) between BonoboBase and a peptide database GenPept. The user wishes to compare a Bonobo microarray experiment, which has BonoboBase
30 numbers, and a human microarray experiment, which has Genbank Accession numbers.

Using some embodiments of the invention: Since GenPept and Genbank may be cross-referenced by some embodiments of the invention, all that may need to be done is add another alias to these records. The user can run a translation table filter

program, and can specify BonoboToGenpept.txt as the input file. Now that the aliases have been added, the user can run a translate file feature as many times as the user wishes to translate the Bonobo microarray experiments to Genbank numbers.

Using a link-federated database: Although the user may get the data file into
5 the database and look at it, automatic translation may not be possible using a link-federated database.

Using a data warehouse: It may be difficult to easily add the new data to the database. The user may have to get a database administrator to create a new set of tables for BonoboBase records, which may be joined to the table of GenPept records.
10 Because there is no grouping, a custom script may then have to be written for this specific type (BonoboBase to Genbank, through GenPept) of translation.

Example 2 - New Experiment

The user decides to screen compounds against some bonobo genes. The user
15 devises a system wherein the user can label each gene-compound interaction with either 'effect' or 'no effect'. When the user acquired the database, the user didn't anticipate performing compound screening, and didn't ask for this feature. Now the user wants to search the database for all the genes in the kinase family that are affected by ethanol.

20 Using some embodiments of the invention: If the user's file is in tab-delimited (for example, an Excel text file) format, in XML format, or in any other format it can import, programming or data structure modification may not need to be done. The user can then search for genes in the kinase family affected by ethanol.

Using a link-federated database: The data can be added by creating a new
25 template for the new format. However, complex queries such as this one may not be possible in a link-federated database because connections generally are hyperlinks and may not be usable in searches.

Using a data warehouse: Again, it may be difficult getting the data into the database. Since the user did not request support for this particular type of data in the
30 beginning, the database structure may need to be modified to add the data. Once this has been done, searches such as the one described can be performed.

Example 3 - Unsupervised Explanation

The user takes a treatment series experiment and uses hierarchical clustering to arrange the data. The user looks at the genes and identifies a sub-tree containing genes that are all decreasing in expression over time in a highly correlated manner.

- 5 Now that the user has a list of genes, the user wants to know why they would be clustered together in this experiment.

Using some embodiments of the invention: The data in the entity-relationship model can be typeless, so one can search for shared groupings of any type with a single query. Using a query tool, the user can enter the gene names, and may be given
10 a result such as "80% of these genes are in the Prosite family EF-hand".

Using a link-federated database: Such queries may not be possible in this type of database. The user may enter the names of the genes one by one, look at the records, write down the families/references/etc. and look over it manually to determine if they had anything in common.

- 15 Using a data warehouse: Since a data warehouse is based around specific tables for specific data types, a typeless grouping search may not be able to be performed. The closest approximation may be a supervised approach, where the user may phrase the question as "What Prosite grouping do these genes share?" Since there are hundreds of possible types of groupings, asking this question for every single one
20 may be extremely tedious.

Example 4 - Distant Relationship

- The user conducts an experiment, which leads the user to believe that Protein CSR2_RAT is connected to Leukemia. The user cannot, however, find any literature
25 or references to confirm this, and wants to search the database for any possible indirect links between CSR2_RAT and Leukemia.

Using some embodiments of the invention: The user can use a relationship finder tool and enter the *CSR2_RAT* and *Leukemia*. Some embodiments of the invention can perform a breadth-first search, traversing any kind of relationship and
30 can tell the user that "CSR2_RAT shares Pfam: LIM with RHM1_HUMAN. RHM1_HUMAN is related to OMIM-DISEASE: Leukemia".

Using a link-federated database: Once again, the task of searching the database for a connection may become a tedious process of clicking between pages,

hoping to find some relationship. It may be difficult to do this automatically, except perhaps using a Web crawler.

Using a data warehouse: As with the previous Example it may be very difficult to perform an unsupervised traversal of the data because it generally is contained in tables of specific types with specific relationships. While the user can ask "Does CSR2_RAT share a Pfam domain with a protein related to Leukemia?", the user may not be able to simply say, "Find the relationship." This may make the search extremely tedious, and it may be virtually impossible if there are more than two steps involved.

10

Example 5 - Multivariable Cluster Analysis

The user would like to look at the hierarchically clustered expression data and understand how the clusters relate to molecular function in the Gene Ontology.

Using some embodiments of the invention: The user can enter clustered expression data and select Molecular Function as a second view. The user then may get a display showing the expression-clustered data in one panel and the same genes as are in this experiment clustered by molecular function in another panel. When the user moves the mouse over a subtree in one panel the genes in the subtree may be highlighted in both panels so that the user can explore and make hypothesis about the relationships between function and expression in the experiments.

20

Using a link-federated database: It may be possible that a program could be written to retrieve every gene record specified in the user's file and the group them by common references. However this may require that there were no levels of indirection (i.e., the gene records directly reference by what they are to be clustered), which is not the case in the Gene Ontology, and that the structure of the tree was flat (i.e., not a hierarchy or ontology).

25

Using a data warehouse: This may be possible, if the data warehouse was designed to support all the levels of the ontology data.

The above Examples illustrate that embodiments of the invention can provide translation among naming systems, allowing cross-referencing and clustering of experimental and/or public data. Data types that have never been seen before can be added. Aliasing and grouping can reduce multiple levels of indirection to a single reference. Complex queries may be performed and typeless data may be used.

30

It also will be understood that although embodiments of the invention have been described above with respect to genes, proteins, literature references, domains, ontologies and other data types, the ways in which data can be categorized and cross-referenced using embodiments of the invention can be virtually unlimited. For example, the description lines of genes from Hugo may be used in order to group them into sets of mutant alleles. A combination of Medline and expression data may be used to infer groupings on the basis of likely interactions. Also, high-throughput screening data may be used to cross-reference chemicals to genes and then group the chemicals by structure. Many other databases also can be used.

The application space for embodiments of the invention also appears to be varied and widely unexplored. Embodiments of the invention can allow a user to perform searches and analyses that previously may have been unavailable or at least very difficult to implement. There are many more applications beyond those described here. Embodiments of the invention can include both remote and local APIs with many powerful functions, both for internal use and to encourage development of applications.

Figure 9 is a flowchart of operations for integrating biological/chemical databases according to other embodiments of the present invention. As will be described below, these embodiments can create an ontology network from a plurality of independent ontologies, to thereby provide a foundation for discovery.

In particular, referring to Figure 9 at Block 902, an entity-relationship model is obtained for each of the plurality of biological/chemical databases. It will be understood that the entity-relationship model may be available as part of the database schema of each of the biological/chemical databases so that it merely may need be received. If not, an entity-relationship model may be created using known techniques. Accordingly, the word obtain, as used herein, includes receiving an existing entity-relationship model and/or creating an entity-relationship model.

Then at Block 904, at least some of the related entities in the entity-relationship models in at least two of the biological/chemical databases are identified. At Block 906, the related identities in the entity-relationship models in the at least two of the biological/chemical databases are linked, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases and creates an ontology network. Operations at Blocks 904 and 906 are repeated until a plurality of related entities, and in some embodiments all related entities, are

identified and linked. Once the ontology network is created, a query may be performed by performing operations of Blocks 512-520, as were already described. This description will not be repeated for the sake of brevity.

In some embodiments of the invention, the related identifies are identical entities that are linked by merging into a single identity. In other embodiments, the related identities need not be identical. In particular, in some embodiments, entities which are similar but not identical may be associated with one another through a relationship type. The two entities may share aliases, inherit relationships from one another, and may share all benefits of a merge, but may remain separate entities. In other embodiments, entities which are similar but not identical may be associated with one another through a parent entity. All of the identical information may be contained in the parent entity in these embodiments, while the differential information is contained in the child entities. Common relationships are inherited through the parent entity, while relationships particular to the child entities are not. Finally, in still other embodiments, entities which are deemed to be related through traversal may be associated through the construction of a meta-relationship which encapsulates the multiple relationships along the original traversal. Yet other examples of linking of related entities may be provided, according to other embodiments of the invention.

Referring now to Figure 10, operations for integrating a new biological/chemical database into a plurality of biological/chemical databases according to some embodiments of the invention now will be described. In particular, as shown at Block 1002, an entity-relationship model is provided for the plurality of biological/chemical databases. The entity-relationship model links at least some related entities in at least two of the biological/chemical databases. This entity-relationship model may be obtained, for example, by performing the operations of Blocks 902-906 of Figure 9.

Still referring to Figure 10, at Block 1004, an entity-relationship model for the new biological/chemical database is obtained. At Block 1006, at least some of the related entities in the entity-relationship model for the new biological/chemical database and the entity-relationship model for plurality of biological/chemical databases are identified. If related entities are identified at Block 1006, the identical entities in the entity-relationship model for the new biological/chemical database and the entity-relationship model for the plurality of biological/chemical databases are linked.

For example, in some embodiments, at Block 1008, the identical entities in the entity-relationship model for the new biological/chemical database and the entity-relationship model for the plurality of biological/chemical databases are merged into a single entity. Also, in some embodiments, at Block 1010, a plurality of aliases are established for the entity that is merged, a respective one of which points to a respective one of the identical identifies in the entity-relationship models in the at least two of the biological/chemical databases. The identification of related entities, merging and establishing of aliases (Blocks 1006, 1008 and 1010, respectively) are continued, until a plurality, and in some embodiments all, related entities have been identified and linked. Operations for deleting records also may be performed at Block 612-618 as was described above.

Referring now to Figure 11, a plurality of biological/chemical databases may be queried according to some embodiments of the present invention, by providing an ontology network that links at least some related entities in at least two of the biological/chemical databases at Block 1102. This ontology network may be provided by performing the operations of Figures 9 and/or 10. Querying may be performed by performing the operations of Blocks 512-520. These operations will not be described again for the sake of brevity.

Additional qualitative discussion of creation of an ontology network according to some embodiments of the present invention now will be provided. Some embodiments of the invention can overlay/merge/associate ontologies and provide extensive cross referencing to other existing data bases, data tables, data repositories, and ontologies. According to some embodiments of the invention, the resulting knowledge layer can provide an ontology network where multiple ontologies and various entities have been linked. The ontology network can bridge previously disparate data repositories, bringing structure to a previously amorphous assembly of independent ontologies of entities and relationships.

According to some embodiments of the invention, this ontology network can provide multidirectional characteristics of parent-child relationships. Specifically, the relationships that hold among the objects or entities of an ontology network can be said to have a character where each entity may have another entity from which it was derived or have or is assigned hierarchical characteristics with regard to another entity. However, since an ontology network need not be limited to this form, other new relationships or hierarchies can be created by the process of overlay, merge

and/or association of entities from other ontologies of interest. This conceptualization of knowledge may be constructed of knowledge from objects of similar domain and can serve as a specification mechanism for the development of a mesh belief system that can deliver experimental insight. This system may provide for the ability to
5 traverse and thereby establish a linked path of relationships creating associations between characteristically unlike entities and also may provide for the revelation of new information and knowledge. The resulting lattice of semantically rich metadata can form an ontology network that can capture the knowledge from the data sources it supports.

10 According to some embodiments of the invention, an ontology network 210 can reside as a part of an information stack related to the basic scientific experiments where enormous quantities of data are collected, for example as was shown in Figure 2. In some embodiments, the ontology network can be located above a conventional integration tool or layer 206 and can provide a knowledge mining tool or layer 110
15 that can be available for hypothesis or question-driven mining as opposed to complex data mining queries typical of data mining applications. Some embodiments of the ontology network can comprise a meta database of terms, entities and/or data relationships that can provide for a more efficient and intelligent analysis of accumulated data.

20 According to other embodiments of the invention, implementation of virtual experiments 112 and discovery 212 that employ this ontology network can provide inference engines. As is well known, the components of an expert system are a knowledge base, which may be implemented according to embodiments of the invention by an ontology network 210, and an inference engine which performs
25 reasoning. According to some embodiments, an inference engine or reasoning software application searches and creates rules by determined pattern matching and then establishes new rules and develops forward chaining of rules. Virtual experiments 112 within the subject field of inquiry can be executed which can significantly enhance accuracies and/or have abilities to correlate observations to
30 original predictive behavior with a broader input of related information than previously may be employed.

Inference engines can be made more accurate as a result of the type designation of relationship, building of newly determined relationships, along with the quantification of the confidence and/or validity assigned to these relationships. As

will be described below, some embodiments of the invention can assign confidence to different traversals and/or variations in selected paths as they are determined or discovered. This characteristic of an ontology network according to some embodiments of the invention can be further integrated into use by the creator of the virtual experiment to add greater value and relevance to data across the broad span of information among the many domains made available in this semantically rich metadata layer.

As was described above, according to some embodiments of the present invention, an ontology network is created by merging, overlaying and/or linking identical objects and/or establishing a relationship between objects/entities in different ontologies. Figures 12-17 conceptually illustrate an example of the creation of an ontology network according to some embodiments of the present invention.

In particular, Figure 12 depicts an ontology that is linked to data fields known to relate to molecular function. Thus, Figure 12 depicts a molecular function ontology 1210. One specific example of such an ontology is the GO Consortium function ontology. In this ontology, relevant data exists where the gene sequence ID 1220 or the protein ID 1230 encoded by the gene sequence has a known function in some physical location 1240 and/or in a particular tissue 1250. The gene sequence ID 1220 also may be linked to raw sequence data 1260 in the molecular function ontology 1210.

Figure 13 illustrates a biological process ontology 1310 which also links to a gene sequence ID 1320, a physical location 1340, a tissue 1350, raw sequence data 1360 and a protein ID 1330. Figure 14 illustrates a cellular component ontology 1410, which also links to a gene sequence ID 1420, a physical location 1440, a tissue 1450, a protein ID 1430 and raw sequence data 1460.

Figure 15 illustrates the linking of the multiple ontologies of Figures 12, 13 and 14 into an ontology network by identifying an identical entity gene sequence ID 1520 and using the identical gene sequence IDs 1220 of the molecular function ontology, 1320 of the biological process ontology and 1420 of the cellular component ontology, to link the molecular function ontology 1210, the cellular component ontology 1410 and the biological process ontology 1310 into an ontology network by reference to the gene sequence ID. A specific example of the linking of Figure 15 may include the three separate GO consortium ontologies and a linkage via SWISS-PROT database entries according to some embodiments of the present invention.

Operations of Figure 9 may be used in some embodiments to link these disparate ontologies.

Figure 16 illustrates an example of another ontology 1610 for protein function, including a protein ID 1630, a gene sequence ID 1620, a physical location 1640, a tissue 1650 and raw sequence data 1660. Figure 17 illustrates adding the ontology 1610 of Figure 16 using the gene sequence ID entity 1720, for example using operations of Figure 10.

As was described above, an ontology can be thought of as a knowledge construct that contains therewithin an answer to a question or a set of beliefs particular to a given domain. Thus, in the example of Figures 12-17, ontologies about biological processes may aid in the determination of what protein might play a role in a particular process. The combination of ontologies results in the creation of an ontology network in Figures 15 and 17, which can yield answers to questions that were not originally expressed by any of the original ontologies as conceived. Thus, an ontology used to express a belief about system A, and an ontology used to express a belief about system B can be associated together according to embodiments of the present invention, to express belief about systems A and B, but to also answer a new query C.

For example, Figure 18 illustrates a query 1810 that can be run by traversing the ontology network of Figure 17. The query can reflect a belief that, for example, nucleic membrane genes are more likely to create protein kinases than anything else. By traversing the ontology network of Figure 17, the cellular component ontology 1410 can reveal which are the nucleic membrane genes, and the molecular function ontology 1210 can reveal which are protein kinases. Since these two ontologies are now linked in an ontology network, an answer to the query may be provided. Thus, an ontology network according to some embodiments of the invention can allow a user to form hypotheses about the role of function in process, or of process in function. Many other hypotheses may be formed.

It will be understood by those having skill in the art that Figures 12-18 illustrate a relatively simple example of linking of ontologies to provide an ontology network. An example of the complexity of linkages that may be available according to some embodiments of the invention is illustrated in Figure 20. The intensity of the implied web created by this network of linkages can continue to develop. The development of density may result in yielding and revealing accurate and relevant

knowledge to accelerate the organization of knowledge. Increased density of relationships between entities, data structures, and ontologies may result in the acceleration of knowledge and the discovery process.

In particular, Figure 19 illustrates an ontology network comparing the
5 Stanford GO Cell Component Ontology and the Stanford GO Biological Process Ontology. In Figure 19, the Stanford GO Cell Component ontology references the same proteins as the Stanford GO Biological Process Ontology, allowing the traversal from structure to function that is shown in Figure 19.

Figure 20 is presented as an example of the linkages displayed in Figure 19
10 and the organization and resulting increased perspective that may be provided by some embodiments of the invention to reveal relevant information surrounding one entity. Some embodiments of the invention can reorder these cross-references in a manner that may enable the mining of vast amounts of information, literally files of data, quickly and easily, without the need for a deep understanding of any of the
15 databases that are included, or of the complex data-mining techniques applied in the back-end. Users may interact with a logically crafted front-end (interface) that provides access to the complete ontology network, without overwhelming users with complex technical queries.

Figure 21 illustrates another example that uses aliases to provide a network of
20 ontologies according to some embodiments of the present invention. In the case of the heredity breast cancer gene, the multiple aliases of related protein and sequence that encodes it, is shown, and a resulting browser view of the gene, protein and sequence is also shown. The browser is an exemplary query tool of the ontology, and can display the many links and alias examples created in the construction of Figure 20
25 in a potentially easy to understand and intuitive view. Thus, in some embodiments of the invention, the power of the ontology can hide the vast knowledge that is stored in its relationships and constructs.

Figure 22 is a block diagram of a data processing architecture that may be used with some embodiments of the present invention. In particular, the construction
30 of expert systems has been the subject of research in computer science. The creation of a knowledge layer, where a significant responsibility beyond simple reasoning is applied to the inference engine, may need to use supercomputing capabilities. In creating ontology networks according to some embodiments of the present invention, it may be desirable to access significant computing resources. The quantity and time

to complete the construction of such an ontology network may be tied to the volume of data in the repositories to be supported by the ontology network and the available computer resources applied during the construction of the metadata referencing the data repositories. Resources ranging from about 30 – 50 gigaflops may be employed
5 in some embodiments, to construct an ontology network in a reasonable time, such as days. Resources ranging up to about 100 gigaflops or more may be used in some embodiments to construct an ontology network to support larger repositories. A computational system able to support more than 100 Gigaflops of computer power may be among the top 500 supercomputers presently available.

10 In some embodiments, the creation and/or execution of the ontology network may use peer-to-peer or grid computing technology. Here, processing cycles from many computers on a network are harnessed, and the application used to create the ontology network may be "gridified" to make the best use of these resources. The construction of such a knowledge layer may be well suited to distribution of the
15 millions of small processes. As a result of increasing efficiencies and decreasing costs to employ computer resources as a grid, the construction of such a meta database that captures the information content of the underlying repositories may become a common part of the mining of complex and disparate data systems. The design and operation of peer-to-peer computing systems are well known to those of
20 skill in the art and need not be described further herein.

An example of a database schema which can be used in an ontology network engine, such as an ontology network engine 300 of Figure 3 or 400 of Figure 4, to store metadata concerning diverse databases in a metadata database such as the metadata database 308 of Figure 3 or 408 of Figure 4, now will be described. It has
25 been found, according to some embodiments of the invention, that the metadata can be stored in a generic database using a conceptual schema that can be implemented using conventional relational database management systems, such as Oracle, MySQL and/or Access.

It will be understood by those having skill in the art that database design may
30 refer to a conceptual schema that exists between the external perception of data (often referred to as an external schema) and the internal on-disk view of data (often referred to as an internal schema). This three-schema architecture conceptualization can enable a programmer to abstract and create various external views of data from the internal view. The conceptual schema can be a composite of all external schemas,

such as the use of tables and columns in a spreadsheet, so that external views can be derived from the conceptual schema, while providing the translation for data recording to the physical schema or on-disk structure.

Referring now to Figure 23, according to some embodiments of the invention, a conceptual schema for an ontology network can itself be embodied as an entity-relationship model. In Figure 23, the individual boxes may represent tables in a MySQL database. These tables are logical groupings of related data. The lines between the boxes represent relationships between common information or cross-references between distinct tables. The entries inside each box represent unique keys or columns of data for each piece of data held by that table or piece of data.

In particular, referring to Figure 23, the boxes enclosed by dashed Block 2310 may be used to define entities including the entity name, entity category, attributes or properties of the entity, and aliases of the entities. The boxes enclosed in dashed Blocks 2320a and 2320b may be used to define relationships, including an identification of the relationship, the attributes or properties of the relationship, and the type of the relationship. The boxes enclosed by dashed Block 2330 define user interface aspects including security aspects. The boxes enclosed by dashed Block 2340 define Uniform Resource Locators (URLs) for external databases that may be used with an entity browser. The boxes enclosed by dashed Block 2350 provide functionality for updating the ontology when a new version of a database is input. Finally, the box enclosed by dashed Block 2360 defines the applications that can be used with an ontology network. It will be understood that at database schema of Figure 23 may be used by those having skill in the art to create a relational database using a conventional database management tool.

Thus, the database schema of Figure 23 is itself represented by an entity-relationship data model. The entities may hold information and may stand alone, or may have relationships between other entities holding data. Thus, the conceptual schema of Figure 23 illustrates the existing relationships that are declared as being true for the data before discovery of new relationships via inference and/or results are presented. This conceptual schema may be used to create a relational database that can provide a network of ontologies according to some embodiments of the present invention.

Referring now to Figure 24, operations for integrating biological/chemical databases and integrating new biological/chemical databases according to other

embodiments of the present invention now will be described. These embodiments assume that database records are provided via XML text records. The use of XML text records and the conversion of non-XML records to XML records are well known to those having skill in the art and need not be described further herein. Moreover, it is assumed that the loader, such as the loader 302 of Figure 3, that is used to load the XML text records also has knowledge of the ontology's semantics based upon the ontology's external data files. As was described above with respect to Figure 23, the ontology semantics also may be extracted from an external biological/chemical database, if they are not already known. Accordingly, *a priori* knowledge of the ontology's entities and relationships is known at the time of loading.

Referring now to Figure 24, operations begin with an XML description of an entity in a biological/chemical database at Block 2402. At Block 2404, the XML description is read. At Block 2406, a list of aliases is obtained from the XML description. At Block 2408, a test is made as to whether an entity with one of these aliases already exists in the network of ontologies. If yes, the existing entity is obtained at Block 2412. If no, at Block 2414, a new entity is created. Source information then is obtained from the XML text at Block 2416.

Continuing with the description of Figure 24, operations for adding the aliases from the XML input to the entity and merging the entity with other entities when the aliases match now will be described. In particular, for each alias in the XML text file (Block 2418), the alias and the source information are added to the entity at Block 2422. At Block 2424, a test is made as to whether the alias exists in another entity. If yes, the other entity is merged with this one at Block 2426. A test is then made at Block 2428 as to whether any aliases remain and, if so, the operations of Blocks 2418-2426 are repeated until none remain.

Operations continue at Figure 25. At Block 2502, parent relationships and associated source information are added to the entity and at Block 2504, parent relationships that no longer exist are removed from the entity. At Block 2506, child relationships and associated source information are added to the entity and at Block 2508, child relationships that no longer exist are removed from the entity. At Block 2512, the attributes are added or updated to the entity.

Still continuing with the description of Figure 25, operations to remove aliases from the existing entity that no longer appear in the XML input now will be described. In particular, for each alias in the entity (Block 2518), a test is made as to

whether this alias exists in the XML text file at Block 2522. If not, the alias is deleted from the entity at Block 2524. Moreover, as a result of deleting the alias from the entity, a test is made at Block 2526 as to whether the entity needs to be split due to the alias deletion and, if so, the entity is split at Block 2528. The operations of Blocks 2518-2528 are completed until there are no aliases left at Block 2532, whereupon operations end.

Accordingly, Figures 24 and 25 illustrate operations for inputting data into the ontology network via an XML text record according to some embodiments of the present invention. During these operations, new entities are constructed and merged, to achieve linking and merging of previously disparate entities. The addition of an ontology may be executed in the same manner. In particular, elements of the ontology are read and operations of Figures 24 and 25 are followed.

For the purpose of loading an ontology into a preexisting network of ontologies, care may need to be taken because entities within the new ontology may have relationships pointing to other entities within the ontology network, and may also have relationships to entities already existing in the ontology network. The operations that were described above in connection with Figure 25 can maintain consistency. Thus, Figure 25 provides embodiments of operations for building new or adding parent and/or child relationships. Removing aliases that may become out of date as a result of an update process also was described. Other new types of relationships, such as reaction right or reaction left or reaction forward or reaction back also may be added, to provide an ability to filter by step.

The following Table describes algorithms that may be used according to some embodiments of the invention, to add an entity and add a relationship using the database schema of Figure 23 and the operations of Figures 24 and 25:

Table

Adding an Entity

Overview

Add the entity information.
 Add an updateInfo for the entity from the external data source.
 Why updateInfos: to differentiate data from different external data sources in order to handle data inconsistency between those sources.

Once in the system, information cannot be deleted until all external data sources that put it there agree that it no longer exists.

UpdateInfos are associated with aliases and relationships.

Add Aliases to the entity.

- 5 The updateInfo is used when adding aliases.

Add the Entity Information.

Algorithm

Add this entity's category to the category table if it is not already there.

Add this entity's information to the entity table.

- 10 Add this entity's attribute information to the entity property table.

Modified Tables

IcCategoryList

New row added with the entity's category if the category doesn't already exist.

IcEntity

- 15 *New row added with the entity's information.*

IcEntityProperty

New row(s) added with the entity's attribute information.

Add an UpdateInfo for the Entity from the External Data Source.

Algorithm

- 20 If the updateInfo is already in the updateInfo table, update its date information.

Otherwise, add the updateInfo information to the updateInfo table.

Modified Tables

IcUpdateInfo

New row added with the updateInfo's information.

- 25 *mLastUpdated column updated with the date information if the updateInfo is already in the table.*

Add Aliases to the Entity

Algorithm

- 30 If the alias is already in the database attached to another entity, then merge that entity with this alias's entity.

This involves taking all the data for the two entities pointed to by the alias and putting it on a single entity, then removing the other entity from the system.

Otherwise add the alias's information to the Alias table.

Associate the specified updateInfo with the alias.

Modified Tables

IcAlias

New row added with the alias's information.

IcAliasUpdateInfo

5 *New row added to associate the updateInfo with this alias.*

IcTypeList

New row added with the alias's type if the type doesn't already exist.

Modified Tables Due To Merging Entities

IcAlias

10 *IcEntityID column changed to point the alias to the merged entity.*

IcEntity

Existing row for the old entity deleted.

IcEntityProperty

Existing row(s) for the old entity attributes deleted.

15 *IcEntityID column updated to point to the merged entity.*

IcRelationship

Existing row(s) for relationships on the old entity deleted.

ParentIcEntityID column updated to point to the merged entity.

ChildIcEntityID column updated to point to the merged entity.

20 *IcRelationshipProperty*

Existing row(s) for attributes on relationships on the old entity deleted.

IcRelationshipUpdateInfo

Existing row(s) for updateInfos on relationships on the old entity deleted.

IcRelationshipID column updated to point to the merged entity.

25 *IcUpdateInfo*

IcEntityID column updated to point to the merged entity.

Adding a Relationship

Overview

Add the Relationship.

30 *A relationship is added between two already-existing entities.*

One entity is the parent, the other is the child.

Each relationship has an associated UpdateInfo for the external data source.

Add the Relationship.

Algorithm

If a relationship of this type already exists between the parent and child, update that relationship's information.

Otherwise add the relationship's information to the relationship table and its attributes to the relationship attribute table.

- 5 Associate the specified updateInfo with the relationship.

Modified Tables

IcRelationship

New row added with the relationship's information.

IcRelationshipProperty

- 10 *New row(s) added with the relationship's attribute information.*

IcRelTypeList

New row added with the alias's type if the type does not already exist.

IcRelationshipUpdateInfo

New row added to associate the updateInfo with this relationship.

15

- Querying of ontology networks according to other embodiments of the present invention now will be described. In particular, Figures 5, 7, 9 and 11 described embodiments for querying the ontology network according to some embodiments of the present invention. However, it will be understood that ontology networks
- 20 according to some embodiments of the present invention can provide a large number of associations among a large number of entities in diverse ontologies. In some embodiments, discovery may take place by querying the ontology network to traverse the ontology network from one entity to another. Stated differently, in some embodiments, a starting entity and an ending entity may be specified, and the query
- 25 results can provide some or all of the paths that can link the starting entity to the ending entity, to thereby obtain new discovery.

- Unfortunately, due to the large number of linkages between entities that may be provided when building real-world ontology networks, the number of paths which link a starting entity to an ending entity may be inordinately large. In these situations,
- 30 it may be difficult to obtain discovery by merely traversing the entities, as was described, for example, in Block 516, due to the large volume of related entities and relationships that may be obtained. However, as will now be described, some embodiments of the invention can provide predefined path rules (Block 324 of Figure

3) and/or user-defined path rules (Block 322 of Figure 3), and allow traversing the ontology network using these path rules as was described at Blocks 514-520.

More specifically, path rules can specify a type of path to traverse, in response to a given type of query. For example, a path rule may specify a specific type of traversal and a specific type of end point for a specific type of starting point. The path rules can be relatively simple, as was described above, but also can be more complex, involving iterations and/or branching. These path rules can, in effect, create new ontologies within the ontology network based on the belief system of the creator(s) of the predefined or user-defined path rules. *A posteriori* knowledge of the relationship between the disparate ontologies may be built into the path rules that are developed to traverse the ontology network. Path rules may be devised with specific semantics in mind based on the data loaded into the ontology network. Thus, the relationships generated when a path rule is applied to a specific starting entity can have a well defined meaning.

Figure 26 illustrates operations that may be performed to traverse the entities in an ontology network using path rules, according to some embodiments of the present invention, as was generally described at Block 518. In particular, referring to Figure 26, at Block 2610, a path rule is obtained either by a user defining a path rule (Block 322), or by obtaining a predefined path rule (Block 324). At Block 2620, the path rule is applied to a specified start point. At Block 2630, the end point or end points found by the path rule are obtained. At Block 2640 a test is made as to whether additional start points are present. If not, at Block 2650, the results of the query may be provided.

Moreover, as also shown in Block 2650, in other embodiments, the start points and end points that are now linked by the path rule can be used to define a new ontology, and can be stored in the metadata database to become a permanent part of the ontology network based upon the belief of the user of the ontology network, rather than merely being a temporary result of a query. In particular, at each step of the traversal through the entities that comprise an ontology network, decisions are made regarding which relationship is selected. Thus, the establishment of a belief at each step or traversal of the system begins to establish multiple steps of order. A decision regarding which step is next in a traversal may be implemented, according to embodiments of the present invention, by providing filtering in the path rules, to thereby create an overall path rule.

Moreover, once a new relationship is declared that is comprised of other steps in the traversal, these rules can be applied by the external schema. Alternatively, they can be physically applied to the internal schema. In other embodiments, a path rule need not persist or be part of the internal schema. Rather, knowledge mining only
5 may need to enable the presentation of this order to the user's results of a study.

At the point of validation of a path, results may yield significant knowledge regarding an entire system of knowledge that is now resident in an ontology network. Thus, with the application of filtering in the path, execution of path rules and/or global filtering according to some embodiments of the present invention, an ontology
10 network can become more than an amorphous set of entities and relationships, and can become more of a rich knowledge base with inherent discoveries therein.

Accordingly, some embodiments of the invention store the query results that are based on the entity-relationship model of the plurality of biological/chemical databases as at least one new relationship in the entity-relationship model, to thereby
15 store knowledge that was derived from the query in the entity-relationship model of the plurality of biological/chemical databases. The ontology network, therefore, can expand based on the knowledge that was obtained as a result of querying the ontology network. In other embodiments, these query results are not stored, so that the query results are not used to modify the ontology network itself.

20 Filtering according to some embodiments of the invention may specify a relationship type, such as part of, derived from, forward reaction or reverse reaction. Filtering according to other embodiments of the invention also can include or exclude specific types of entities, such as symbols or reactions. Filtering according to yet other embodiments of the invention may also filter on a relationship attribute, entity
25 attribute, alias type, alias ID, category, relationship-type confidence, parent-child, self, and/or other characteristics. Thus, filtering on each step of the traversal can create a preselected path that is acceptable or unacceptable relative to the confidence of the relationship, or as simple as the direction of reaction catalyzed by an agent.

Figure 27 provides an example of an *in silico* experiment that can be derived
30 from an ontology network according to some embodiments of the invention. The example in Figure 27 begins with an experiment 2702, such as two GenBank IDs that both express in an expression data experiment. The remaining blocks of Figure 27 illustrate a path route taken from the starting GenBank ID to the ending GenBank ID. Running the experiment in an ontology network according to some embodiments of

the present invention can validate the path. Moreover, repetition of the path illustrated in Figure 27 across the entire contents of the ontology can implement long-range order in the ontology network and create knowledge and/or values of many other GenBank IDs. A path, such as a path described in Figure 27, can be
5 incorporated into the ontology network, so as to allow this path and all related paths to persist. This can add another ontology to the ontology network according to some embodiments of the invention. Alternatively, in other embodiments this path can be recognized as part of the external schema, and reported as a query result. In either case, a single verified and validated segment of knowledge can be multiplied by
10 inference, and can yield answers to questions or experiments not yet run.

Figures 28-35 provide another example of a path rule that may be used to obtain discovery according to some embodiments of the present invention. In particular, Figure 28 illustrates a small portion of an entity-relationship model that is part of an ontology network according to some embodiments of the present invention.
15 As shown in Figure 29, this example of a path rule can start with a general protein function 2910, and can find the proteins with that function (Block 3010 of Figure 30). The path rule then can expand the query by finding the processes in which the protein is involved, as shown at Block 3110 of Figure 31. All the proteins in these processes may be examined, as shown at Blocks 3210 and 3220 of Figure 32. Screening data
20 can be traversed for the proteins, as shown at Blocks 3310 and 3320 of Figure 33. A list of chemicals that screen favorably can be retrieved, as shown at Blocks 3410, 3420, 3430 and 3440 of Figure 34. Finally, as shown at Blocks 3510 and 3520 of Figure 35, those chemicals with undesirable properties, such as toxicity and/or unwanted structure, can be filtered out.

25 Figure 36 illustrates an example of a user display screen that may be used to initiate a query using the path rule that was specified in Figures 28-35. Figure 37 illustrates a user display screen of query results that may be obtained.

Figures 38 and 39 are flowcharts of operations for querying an ontology network according to other embodiments of the present invention. Figure 38
30 illustrates querying from a user perspective. Figure 39 illustrates operations from a client-server standpoint.

According to other embodiments of the present invention, an ontology network can be constructed where the relationships between objects are further labeled and characterized with confidence levels as well as type. The ontology

network may be traversed in response to a query, to thereby obtain query results that are based on the entity-relationship model including the at least one confidence level that is assigned. Inferences and correlations commonly employed in the biotechnology area may be characterized to better enable application of these relationships as a more exact and analytical science. This knowledge may not only be harnessed by reasoning engines to create more valid and accurate virtual experiments, but also new relationships may be discovered, built into the ontology network, and/or learned by the ontology network to establish and discover new correlations. The value or quality of these new relationships can be screened and/or further characterized.

In some embodiments of the present invention, information queries of the ontology network can be exact. Results of queries where the retrieved information appears to have been filtered can result from the deployment of knowledge associated with preselected paths. In conventional data queries, data acquired may be filtered to screen unwanted and incorrect results. Not only may this be time consuming, but often the results may still contain significant error and false information. In contrast, queries constructed and run using preselected paths according to some embodiments of the invention may provide only an accurate and concise representation of the information content of the underlying repositories.

In view of the above, some embodiments of the present invention have recognized the principle that relationships between biological entities may be critical to the discovery process. Embodiments of the present invention can logically organize and cross-reference data into groups, so that the data can be fully accessible and useful. Some embodiments of the invention can merge naming conventions or aliases. Other embodiments of the invention can allow researchers to place proprietary research data into the broadest possible relative context with public research data. Moreover, some embodiments of the present invention can anticipate researchers, think, reduce or eliminate repetitive tasks and/or automate the manual processes that may be used in research and discovery.

Some embodiments of the present invention can merge and adjust multiple ontologies to reflect the rapidly changing state of standards and semantics in the life sciences, so that legacy work and investment need not be lost. Thus, some embodiments of the invention can converge information relating to biological and chemical properties, physiology and/or published research. This information may be

cross-referenced. For example, cross-referenced information from more than twenty public life sciences databases, including over forty naming systems, may be provided in some embodiments of the invention, and links may be established between genes, proteins, biochemical pathways, diseases, organisms, literature references and other entities of interest that are referenced in each included data source.

Accordingly, some embodiments of the invention can merge redundant database entries from different sources into single entities with alternate names or identifiers. Relationships between entities can capture knowledge from different data sources. These entities and relationships can make up an emergent ontology-based network, capturing the concepts behind life sciences databases. This network may not be hard-coded, such that new entity types can be added without the need to modify the underlying database, and relationships between any entities may be allowed. In addition, in many embodiments, entities are sparsely populated, so that only aspects of original data that either involve relationships between entities, or are relevant to user queries may need to be integrated.

Some embodiments of the invention can represent data as entities. Some embodiments of the invention can allow entities to represent any concept or type, including concepts not already represented in the existing entity-relationship model. Because of this, a user can add a completely new concept or type without the need to make changes to the underlying database.

An entity can represent a single concept type or individual of that type. According to some embodiments of the invention, if that concept is present in multiple data sources, the multiple sources are merged into a single entity. For example, the predicted *C. elegans* protein YKD3_CAEEL or Q03561 from SWISS-PROT also is represented in PIR as S28280, and in WormPep as B0464.3 or CE00017. In some embodiments of the invention, these database entries can be collapsed into a single entity with the individual identifies as aliases. In practical usage, a user can access all of the relationships for the entity by querying with any of its aliases.

In some embodiments, information about an entity, such as its description, molecule type, or annotation, is stored in attributes. In some embodiments, entities can have unlimited attributes, and each attribute has a type and a value. As with entities, attribute types can represent any concept, and new attribute types can be added without the need to make changes to the underlying database. Attributes may

store information about an entity for the purposes of searching and filtering, and therefore can be metadata storage containers. For example, a nucleotide entity may have both a description attribute and an attribute "molecule type", indicating whether it is DNA, RNA, mRNA, etc., but may not have its nucleotide sequence as an attribute. Instead, the locations of the original database records may be cross-referenced by the nucleotide entity, providing a way to fetch the sequence if need be. Because of this, in some embodiments of the invention, entities may be sparsely populated.

In other embodiments, entities also may be organized into categories or classes, which, like entity types, can be added without the need to change the underlying database. Categories may be used for broad binning of entities, for example protein, pathway, literature or nucleotide-sequence.

Some embodiments of the invention may be constructed from life science databases that have either cross-references to other databases, or lists of alternate names. When a source is imported, entities may be created not only for the source records, but also for the database records they cross-reference. This can be thought of as a virtual database entry. If at a later time that record is loaded, then its information may be added to the entity in some embodiments. In this way, relationships may be built up from multiple sources.

Entity-relationship models according to some embodiments of the invention also can include relationships, which can allow one entity to represent a group of other entities. For example, a set of enzyme entities can be grouped into a pathway entity. The pathway is the parent of the enzymes, and they are the children of the pathway. The enzymes are siblings of each other. Each enzyme is linked to the pathway by a single relationship, and because there is a parent and a child, it is a directional relationship.

In the above example, an enzyme may be grouped into a pathway. In addition, an enzyme may be grouped with other enzymes having the same function, for example in the EC classification ontology. In this way, an entity can be a member of an unlimited number of groups, and each group can represent a different aspect of its members, according to some embodiments of the invention.

Just like entities, relationships can have a type and attributes, in some embodiments of the invention. The type may be used to describe the action of the relationship (i.e., a gene product is transcribed from a gene, or a gene product is

translated to a protein), while attributes can contain information about the relationship, such as annotation or ontological information (for example, is-a or part-of). Entities can be thought of as nouns, while relationships may be thought of as verbs.

5 Some relationships may be more certain than others. For example, an enzyme that is known to bind to a ligand is a high quality relationship. On the other hand, if a gene product is said to be related to a protein based on sequence homology of 30%, then that relationship may be of low quality. Therefore, in some embodiments, relationships may have a confidence value to reflect the quality of either the data
10 source or the method used to specify that relationship. Confidence values allow a user to filter out relationships that are of too low quality for their purpose. Because of the confidence values, embodiments of the invention can also be thought of as a DWG.

 There can be many sources for relationships in life science databases. For
15 example, SWISS-PROT cross-references EMBL and GenBank entries, that code for its proteins. A Unigene entry points to similar proteins and ESTs. Enzyme entries reference all the proteins with the specified function. A KEGG pathway contains a list of enzymes. Medline entries point to MESH headings, as well as to gene, protein and chemical accession numbers. In this way, a complex network of relationships can
20 be built according to embodiments of the invention. For example, a set of relationships can connect an EST to a gene product, which is in turn grouped under a protein, which is classified as an enzyme with a known function, which has known chemical ligands and is grouped in a pathway. The set of entity- and relationship-types that define the steps to go (in this case) from DNA to chemical ligand provide
25 an example of a path.

 The path above starts at a sequence and ends at a chemical ligand while traversing the specified steps in between. Defining this path and traversing it may be a time-consuming lookup task, for example, from a long list of up- or down-regulated genes from a microarray experiment. Manually traversing the path may require
30 looking up entries in multiple databases, from GenBank to Unigene to SWISS-PROT to Enzyme to KEGG and Ligand. Because embodiments of the invention may be a DWG, it can become a graph theoretical operation to automate the process of traversing the path in an efficient manner. In this way, complex cross-referencing tasks may be collapsed into a single operation.

Some embodiments of the invention can use a specification of rules that define paths using XML. A simple rule is a single step, a path rule is multi-stepped, and a branch rule has conditional branching. A full path may contain different combinations of rule types, and a branch or path rule type can have subrules of any type. In addition, each rule can filter by attribute, type or category. The overall specification of a path defines input and output types or categories.

Some embodiments of the invention also can capture ontological relationships implicitly and/or explicitly. In particular, an entity can explicitly represent an ontological concept. In this case, its parents are more general concepts and its children are more specific concepts. A relationship's type defines how a child concept relates to its parent. Concept entities can also represent groups of instances of that concept. In the above example, a DNA polymerase entity constructed from SWISS-PROT has an is-a relationship with the concept entity parent EC:2.7.7.7 (DNA-directed DNA polymerase), and also has a part-of relationship with the parent GO:0006260 (DNA replication). The EC entity has the more general parent EC:2.7.7.- (nucleotidyltransferases), which has the more general parent 2.7.- (transferring phosphorous-containing groups). At the top of the hierarchy rests EC:2.-.-, which is the general classification of transferases. All of the DNA polymerases grouped under the 2.7.7.7 entity are siblings with the same function, while all of the entities group under GO DNA replication are all siblings in the same process.

Some embodiments of the invention also can define an ontology implicitly. In particular, each entity type and category is a concept, while its relationships define the ontological framework. For example, a protein entity is encoded by a group of gene products, each of which is transcribed from a gene. These relationships are built from the cross-references in life science databases. When a new entity type is added, or an entity is put in a relationship with a previously unrelated entity type, new knowledge about how the different entity types relate to each other may be created.

Since an ontology represents a knowledge domain, an entity that has relationships to entities in more than one domain can bridge those domains. In some embodiments, bridge entities are typically experimental or analytical results. One example is the bridging of biology and chemistry, centered around human beta 2 adrenergic receptor (B2AR) and clenbuterol. SWISS-PROT cites two cloning references that show B2AR is expressed in several tissues, including blood and brain, and is classified by GO as being involved in adenylate cyclase activation. The

SWISS-PROT record points to at least 11 nucleotide sequences for the receptor, and it is classified by Prosite, Interpro and Prints as having GPCR domains. At least two articles referring to this protein are linked to asthma MESH headings, and OMIM links B2AR to asthma as well.

5 In the chemical domain, it is known that clenbuterol is also known as planipart and clenbuterolum (ChemIDPlus), and it is used as a bronchodilator (ChemIDPlus). Its structure can be retrieved from ChemIDPlus, which can indicate that the chemical has several functional groups. Fingerprinting analysis can bring up structural similarity to several other drugs, including Albuterol.

10 To bridge the two domains, experimental data may be used. In this case, text mining of the journal *Biochemical Pharmacology* shows a 70nM binding constant Kd between clenbuterol(-) and B2AR. In some embodiments of the invention, the domains can be bridged in at least two ways: an experimental result entity can be created that links chemical and receptor, or a relationship between protein and ligand
15 may be created. A path may then be traversed from ligand to protein to disease, and from ligand to clinical application, which can show that clenbuterol is a bronchodilator used to treat asthma.

 Side effects also may be predicted: adenylate cyclase activation leads to increased protein kinase A activity (CSNDB), which increases the responsiveness of
20 cardiac muscle to calcium currents (CSNDB). Not surprisingly then, clenbuterol increases heart rate and can in some cases cause cardiac arrhythmia (text mining of HSDB).

 Additionally, other structurally similar drugs can be analyzed to anticipate their action. Albuterol, as mentioned above, is structurally similar to clenbuterol.
25 Although there may be no screening data for albuterol, it can be predicted that it is also a beta 2 adrenergic agonist, can be used to treat asthma, and is associated with similar side effects.

 Thus, embodiments of the invention can provide context to high-throughput life-science experiments by improving information retrieval, and by enhancing
30 automation and data mining ability. In some embodiments of the invention, new data is merged with existing data, and the resulting entities capture the knowledge and relationships of both sources. Both relationships and entities can have a type for filtering, and attributes for capturing relevant data from original sources. Because of merging and grouping, the resulting ontology network can be more highly connected

than the original data sources, which can allow a path to be found between entities in previously unrelated knowledge domains. Moreover, once a path is defined by a user, it can be used in high throughput analyses, such as a microarray results annotation pipeline.

5 Additional Examples

The following additional examples shall be regarded as merely illustrative and shall not be construed as limiting the invention. Some embodiments of the present invention can be used to integrate databases other than biological/chemical databases, to create an ontology network. The following additional examples illustrate how

10 three diverse ontologies in the form of databases relating to personal data, securities data and government data can be integrated into an ontology network.

More specifically, referring to Figure 40, one or more databases related to personal data 4010, one or more databases related to securities data 4020 and one or more databases related to government data 4030 can be integrated into an ontology network 210 by obtaining an entity-relationship model for each of the databases 4010-4030, identifying related entities in the entity-relationship models of at least two of the databases 4010-4030, and linking at least two of the related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of databases. The ontology network 210 may be used for discovery, prediction and simulation 212, as was already described, for example, in connection with Figure 2.

Figure 41 illustrates a more detailed example of the linking of related entities in entity-relationship models for a plurality of databases. More specifically, Figure 41 provides a simplified entity-relationship model for a plurality of databases related to personal data 4110, a plurality of databases related to securities data 4120, and a plurality of databases related to government data 4130, which may provide an embodiment of databases 4110–4130, respectively, of Figure 40.

As illustrated in Figure 41, the databases related to government data 4130 may include entities for government statistics that may be published on a regular basis, and that constitute databases of economic indicators that can impact options trading of the ten and thirty year government notes which, in turn, can impact the sales of bonds and mutual fund price shares. In particular, entities for Gross Domestic Product (GDP) 4131, job growth 4132, consumer confidence 4133, weekly retail sales 4134, earnings and growth 4135, and monthly retail sales 4136, are related to an economic indicators entity 4137.

As is well known to those having skill in the art, the data in the GDP entity 4131 is a measure of the nation's total output of goods and services. The data in the job growth entity 4132 provides an indicator of whether the job market is expanding or contracting. The data in the consumer confidence entity 4133 is an index of
5 consumer sentiment based on monthly interviews with 5000 households. Weekly retail sales data in entity 4134 is reported by the Census Bureau. The Census Bureau also reports monthly retail sales data in entity 4136. Data for the earnings growth rates entity 4135 is also reported by the federal government.

The entities 4131-4136 are all related to an economic indicators entity 4137.
10 The economic indicators entity 4137 is linked to a federal discount rate or discount rate futures entity 4140 which also includes a rate history entity 4141 and a guidance entity 4142. The federal discount rate or discount rate futures entity 4140 is in turn linked to a conference board options value of TNX/TYX (options on the ten year and thirty year rate) entity 4143. It will be understood that the government data 4130 that
15 is shown at the right-hand side of Figure 41 represents a simplified entity relationship model of many government databases related to economics.

It also will be understood that government data 4130 generally is tabulated in a number of databases on a large number of related and seemingly unrelated topics. In addition to the entities shown in Figure 41, other examples include the money in
20 circulation, M1, M2 and M3, and many other such financial numbers. In addition, the government tabulates crop data, weather statistics, weather forecasting, geothermal, geographic, interstellar, gravitational and commodities data. While this data may be relevant to commodity, futures and option trading, such as takes place at the Chicago Mercantile Exchange or the CBOE Exchanges, experts can create relationships or
25 postulate theories of relationships between many of these data types and factors, and their eventual impact on securities markets and/or the value of particular stocks, bonds and mutual funds containing financial instruments of related companies. These expert traversals and/or relationships can be captured in some embodiments of the present invention, for exploitation and application by expert users and/or by less
30 expert users.

Still referring to Figure 41, an entity relationship model related to securities data 4120 also may be provided. The entity-relationship model related to securities 4120 may include an entity for stock indexes 4121, an entity for industry indexes 4122, and an entity for industry sectors 4123. These entities in turn relate to a

companies entity 4124. The companies entity 4124 may be related to a corporate bond entity 4127, which in turn can be related to an interest entity 4128 and a current yield entity 4129. A mutual bond fund entity 4125 may be related to a mutual fund shares entity 4126, which in turn can be related to the interest entity 4128 and the
5 current yield entity 4129.

In particular, many databases exist related to stocks 4121, bonds 4127 and mutual funds 4126. Each of these databases may represent an entity type and may be composed of many different company stocks, bonds or fund shares. An example of an extensive database of this type is the Value Line database of stocks. In this
10 example, Value Line has tabulated about 280 financial characteristics or data items of each company in the list. Their list includes about 6000 different companies in different sectors of the economy. These characteristics can include their proprietary characteristics, such as technical rank and safety rank, and general data such as Beta, relative price-to-earnings ratio, earnings-per-share (current and trailing 12 months),
15 stock price (high/low) and 200 or more other factors that are tabulated for each company. Other related and similar data exist for bonds and mutual funds.

Each of these entity types, as well as each type of stock, bond or mutual fund, may exist in one or more indexes, such as bond indexes, stock indexes and mutual fund indexes. Many of these indexes also are tabulated, and have trading vehicles on
20 the American Stock Exchange, the New York Stock Exchange, or NASDAQ. Many of these entities, such as stocks, bonds, mutual funds and indices, are part of or related to an industry segment. These industry segments have related indexes 4122 and trading vehicles as well.

A particular company sells bonds, sells stocks, creates earnings and is part of
25 mutual funds which also creates earnings, dividends and/or interest. Options (securities derivatives of the above instruments) may be impacted by or tightly related to the underlying securities and react accordingly.

Accordingly, an ontology can be created from the above securities data types 4120, and integration or association of ontologies that can result in the creation of an
30 ontology network according to some embodiments of the present invention. These ontologies may be filled with a great deal of information and relationships, and can be tabulated and stored.

Still referring to Figure 19, an entity-relationship model related to personal data 4110 may relate to an individual. In particular, a capital gains entity 4111

identifies capital gains in an individual's portfolio, and a portfolio entity 4112 can include a securities balance and a database of personal preferences.

As also shown in Figure 41, a related entity in at least two of the entity-relationship models is identified. In particular, in Figure 19, the stock index entity 4121 and industry index entity 4122 in the securities data entity-relationship model 4120 are related to the economic indicators entity 4137 of the government data entity-relationship model 4130. Also, the option entity 4143 is related to the corporate bond entity 4127 and mutual fund shares entity 4126. Finally, the capital gains entity 4111 in the personal data entity-relationship model 4110 is related to the mutual fund share entity 4126 and corporate bond entity 4127 of the securities data entity-relationship model 4120. Thus, at least some of the related entities that are identified are linked, to thereby create an entity-relationship model that integrates the plurality of databases.

A more detailed description of how the integrated entity-relationship model of Figure 41 may be used by an individual to make portfolio position modifications now will be described. In particular, as also shown in Figure 41, a path rule may be identified that may link the economic indicators entity 4132 and the portfolio entity 4112 using the relationship path rule 4150 that is shown by bold linking arrows in Figure 41.

As the economic indicators 4137 change, they can have an effect, by directly impacting the federal discount rate via Federal Reserve Board action, and/or by impacting the perceived federal discount rate futures 4140. This economic data and/or federal action, will impact the CBOE options of TNX and TYX 4143, which are options on the ten year and thirty year Treasury bond rate, and are based on the yield to maturity of the most recently auctioned respective treasury bond. Changes in the value of these instruments are widely watched, and the movement or change in its value can impact the current market, both positively and negatively regarding the sale of corporate bonds 4127. These instruments may change the current yield 4129, and/or may result in further change in value as changes occur in the options market for government securities. Bond fund shares 4125 may also change in value and may further sustain changes in current yield, and/or impact value and cause changes in the interest rates assigned to new issues. Finally, these changes can directly result in a capital gain/loss 4111 from the purchase or sale of these equities. This can impact a portfolio database entity 4112, that includes information on personal preferences of a

customer, and an adjustment or rebalancing of a customer portfolio can be recommended.

The above example shows that there can be relationships to a portfolio balance 4112 that can reside not merely in the databases directly associated with the securities data 4120, but that can reach further into information warehouses that are removed 5 from the databases relating to the relevant securities data 4120. This example can be expanded to capture knowledge from those expert in the field that can delineate some or many of the complex relationships that can exist between actions or activities in a global sense that may have a perceived relationship to a portfolio balance, while being 10 remote and/or indirect in a parent/child relationship.

Accordingly, ontology networks, according to some embodiments of the present invention, can be applied to the investment community. In the investment community, investment firms and brokerage houses hire associates to act as portfolio managers or customer client managers. They may have little expert knowledge with 15 regard to the relationships and actions that might indirectly or directly impact particular instruments. Commodity contracts or related security derivatives are examples of such instruments that may be impacted by many peripheral activities or actions that can occur. These actions can include economic, environmental and any other activity, action, event or data that in some way can be related by a combination 20 of traversals to the file, commodity or derivative in question. There presently appears to be a significant need in the securities industry to capture the expert knowledge of the highly experienced investors/traders who may derive their strategies and plans from what could be represented in an ontology network as traversals and association of relationships between key indicators, databases, events, actions and their expected 25 impact on companies and related securities. Embodiments of the invention can allow this expert knowledge to be captured and exploited.

Figure 42 is a more detailed example of an entity-relationship model that integrates a plurality of databases according to some embodiments of the present invention. In Figure 42, relationship types also are indicated. This entity-relationship 30 model may be used to obtain expert advice as to the advisability of a major purchase 4210 based on the integration of government data, securities data and personal data. Accordingly, an ontology network that comprises relationships between securities data from companies and information and relationships contained within a number of government databanks, can be used to create a valuable tool to capture expert

knowledge in this area for use and application by less skilled industry participants and/or by individuals.

Finally, it will be understood that Figures 40-42 provided examples of the integration of personal data, securities data and government data into an ontology network. However, ontology networks may be created in many other fields. Several examples now will be generally described in the fields of criminology/law enforcement, a government budget and the weather. Many other examples may be envisioned by those having skill in the art.

In the field of criminology and law enforcement, data repositories may exist that store retained fingerprint and comparative matching algorithms, DNA data and large databases of information on individuals, where this information on individuals has been generated through either elicit (criminal) activity and/or benign activities, such as public employment. Moreover, local, national and international databases are being developed which include crime scene information and characteristic observations of various crimes. These different ontologies can be merged into an ontology network that could be used, for example, by a task force or other activity whose aim is to understand the nature of organized criminal activity, by integrating the data repositories that are developed on organized crime activities with a host of specific local crime scene information. The relationships that can be established between organized crime activities, national fingerprint databanks, and local crime scene data repositories, can provide an ontology network that can provide new insight into the activities of a criminal organization and/or a clearer focus on their objectives.

In the field of government budgets, it is known that the development of public policy and budgeting for local and national purposes represents a fine balance between the application of funds to various activities relative to public opinion or policies. Accordingly, a relationship may exist between funds that may be available for public welfare, or the creation of new programs, such as a nationally-supported drug subscription plan, and criminal activity on a local, national or international scale. An ontology network, according to some embodiments of the present invention, that integrates international, national and/or local budgetary information and law enforcement data, can be used to provide a predictable understanding of relevant opinion, the results of which may impact other seemingly unrelated programs. This ontology network could be extended to national security, since related data being acquired, as well as the expenses that are entailed, may have an impact on other

totally unrelated expenses, and may also have an impact on public opinion and the resulting policy.

As a final example, an ontology network that uses weather data according to some embodiments of the present invention now will be described. In particular, 5 documentation of world weather patterns can enable the prediction of the character and depth of droughts and heavy rain activity. Other global patterns may be observed with regard to development and progress of storms. These data repositories are being accumulated at significant cost worldwide, and include details and analysis of global data, including data relating to the characteristics of a single storm or weather event, 10 as well as generalizations and characteristics of weather events as types. It is further known that weather events can impact crop yields, with the resulting expectations of profits and losses resulting in impacts to certain related futures trading that may also be occurring on global futures markets. Futures trading and changes in the value of futures contracts can impact the resulting decisions by farmers as to their expectations 15 for profit and planting decisions for the next season. While this may directly impact the general food supply, the futures activities may also impact decisions by farm equipment manufacturers to manufacture farm equipment, which in turn can impact raw materials costs and future buying patterns of commercial buyers in industries related to these material acquisitions. An ontology network according to some 20 embodiments of the present invention can merge ontologies related to weather, crops data, futures trading, farm equipment manufacturing and raw materials. This ontology network then can be traversed by an expert, to establish a path rule for retention of the expert knowledge. Thus, expert thinking can be captured to create a representation that can clearly identify the impact of weather on the cost of steel for 25 increased farm equipment production in the coming year, as an example.

In the drawings and specification, there have been disclosed typical preferred embodiments of the invention and, although specific terms are employed, they are used in a generic and descriptive sense only and not for purposes of limitation, the scope of the invention being set forth in the following claims.

What is Claimed is:

1. A method of integrating a plurality of biological/chemical databases, comprising:
obtaining an entity-relationship model for each of the plurality of biological/chemical databases;
5 identifying related entities in the entity-relationship models of at least two of the biological/chemical databases; and
linking at least two of the related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases.
10
2. A method according to Claim 1 wherein at least one of the plurality of databases represents an ontology and wherein the entity-relationship model that integrates the plurality of biological/chemical databases creates an ontology network.
- 15 3. A method according to Claim 1 wherein the related entities are identical entities and wherein linking comprises merging the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases.
- 20 4. A method according to Claim 3 wherein the merging further comprises establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases, a respective alias of which refers to a respective one of the at least two of the identical entities that are identified.
25
5. A method according to Claim 1 further comprising:
traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of
30 biological/chemical databases.
6. A method according to Claim 5 wherein the traversing comprises:

traversing the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending entity to thereby identify relationships between the starting entity and the ending entity that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

7. A method according to Claim 5 wherein the traversing comprises: traversing the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

8. A method according to Claim 5 wherein the traversing comprises: traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query and in response to at least one path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

9. A method according to Claim 8 wherein the at least one path rule specifies a type of path to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of path not to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of ending entity that can be included in the query results, a type of ending entity that is not to be included in the query results, a type or class of relationship to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type or class of relationship that is not to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases and/or a confidence level to be achieved in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases.

10. A method according to Claim 8 further comprising storing the query and the path rule for reuse.

11. A method according to Claim 5 further comprising:
5 storing the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases as at least one new relationship in the entity-relationship model that integrates the plurality of biological/chemical databases to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of
10 biological/chemical databases.

12. A method according to Claim 5 further comprising:
assigning a confidence level to at least one of the relationships in the entity-relationship model that integrates the plurality of biological/chemical databases.

15
13. A method according to Claim 12 further comprising:
traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of
20 biological/chemical databases including the at least one confidence level that is assigned.

14. A method of integrating a new biological/chemical database with a plurality of biological/chemical databases, comprising:
25 providing an entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in at least two of the biological/chemical databases;
obtaining an entity-relationship model of the new biological/chemical database;
30 identifying related entities in the entity-relationship model of the new biological/chemical database and the entity-relationship model of the plurality of biological/chemical databases; and

linking at least two of the related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

5 15. A method according to Claim 14 wherein the entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in the at least two of the biological/chemical databases provides an ontology network and wherein the entity-relationship model for the new biological/chemical database represents an ontology.

10

 16. A method according to Claim 14 wherein the related entities are identical entities and wherein the linking comprises merging the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases and the new
15 biological/chemical database.

 17. A method according to Claim 16 wherein the merging further comprises establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases and
20 the new biological/chemical database, a respective alias of which refers to a respective one of the at last two of the identical entities that are identified.

 18. A method according to Claim 17 wherein the new biological/chemical database is an updated version of one of the plurality of biological/chemical
25 databases, the method further comprising:

 identifying at least one entity in the one of the plurality of biological/chemical databases that has been deleted from the updated version of the one of the plurality of biological/chemical databases; and

 removing an alias that is associated with the at least one entity that has been
30 removed.

 19. A method according to Claim 18 further comprising:

splitting at least one entity in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database based upon the alias that was removed.

5 20. A method according to Claim 14 further comprising:

identifying entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database; and

10 adding at least one new entity to the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database that corresponds to the entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new
15 biological/chemical database.

21. A method according to Claim 14 further comprising:

traversing the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response
20 to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

22. A method according to Claim 14 further comprising:

25 traversing the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response to a query and in response to at least one path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

30

23. A method according to Claim 21 further comprising:

storing the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database as at least one new relationship in the entity-relationship

model that integrates the plurality of biological/chemical databases and the new biological/chemical database to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

5

24. A method according to Claim 14 further comprising:
maintaining an image of the entity-relationship model of the plurality of biological/chemical databases prior to the linking.

10

25. A method according to Claim 24 further comprising:
comparing the image of the entity-relationship model of the plurality of biological/chemical databases prior to the linking and the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

15

26. A method according to Claim 14 wherein the entity-relationship model of the new biological/chemical database does not include relationships therein.

27. A method of querying a plurality of biological/chemical databases,
20 each of which includes records for a plurality of biological/chemical entities, the method comprising:
providing an integrated entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in at least two of the biological/chemical databases; and
25 traversing the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

30

28. A method according to Claim 27 wherein the traversing comprises:
traversing the integrated entity-relationship model of the plurality of biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending entity to thereby identify

relationships between the starting entity and the ending entity that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

29. A method according to Claim 27 wherein the traversing comprises:
5 traversing the integrated entity-relationship model of the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

10

30. A method according to Claim 27 wherein the traversing comprises:
traversing the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query and in response to at least one path rule to thereby obtain query results that are based on the integrated entity-
15 relationship model of the plurality of biological/chemical databases.

31. A method according to Claim 30 wherein the at least one path rule specifies a type of path to use in traversing through the plurality of entities, a type of path not to use in traversing through the plurality of entities, a type of ending entity
20 that can be included in the query results, a type or class of ending entity that is not to be included in the query results, a type or class of relationship that is to be used in traversing through the plurality of entities, a type of relationship not to be used in traversing through the plurality of entities and/or a confidence level to be achieved in traversing through the plurality of entities.

25

32. A method according to Claim 30 further comprising storing the query and the path rule for reuse.

33. A method according to Claim 27 further comprising:
30 storing the query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases as at least one new relationship in the integrated entity-relationship model of the plurality of biological/chemical databases to thereby store knowledge that was derived from the

query in the integrated entity-relationship model of the plurality of biological/chemical databases.

34. A method according to Claim 27 further comprising:
5 assigning a confidence level to at least one of the relationships in the integrated entity-relationship model of the plurality of biological/chemical databases.

35. A method according to Claim 34 further comprising:
traversing the integrated entity-relationship model of the plurality of
10 biological/chemical databases in response to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases including the at least one confidence level that is assigned.

36. A system for integrating a plurality of biological/chemical databases,
15 comprising:
an entity-relationship model for each of the plurality of biological/chemical databases;
means for identifying related entities in the entity-relationship models of at
20 least two of the biological/chemical databases; and
means for linking at least two of the related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases.

37. A system according to Claim 36 wherein at least one of the plurality of
25 databases represents an ontology and wherein the entity-relationship model that integrates the plurality of biological/chemical databases creates an ontology network.

38. A system according to Claim 36 wherein the related entities are
30 identical entities and wherein the means for linking comprises means for merging the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases.

39. A system according to Claim 38 wherein the means for merging further comprises means for establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases, a respective alias of which refers to a respective one of the at least two of the identical entities that are identified.

40. A system according to Claim 36 further comprising:
means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

41. A system according to Claim 40 wherein the means for traversing comprises:
means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending entity to thereby identify relationships between the starting entity and the ending entity that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

42. A system according to Claim 40 wherein the means for traversing comprises:
means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

43. A system according to Claim 40 wherein the means for traversing comprises:
means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query and in response to at least one

path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

44. A system according to Claim 43 wherein the at least one path rule
5 specifies a type of path to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of path not to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of ending entity that can be included in the query results, a type of ending entity that is not to be included in the query results, a
10 type or class of relationship to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type or class of relationship that is not to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases and/or a confidence level to be achieved in traversing through the entity-relationship model that integrates the
15 plurality of biological/chemical databases.

45. A system according to Claim 43 further comprising means for storing the query and the path rule for reuse.

20 46. A system according to Claim 40 further comprising:
means for storing the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases as at least one new relationship in the entity-relationship model that integrates the plurality of biological/chemical databases to thereby store knowledge that was derived from the
25 query in the entity-relationship model that integrates the plurality of biological/chemical databases.

47. A system according to Claim 40 further comprising:
means for assigning a confidence level to at least one of the relationships in
30 the entity-relationship model that integrates the plurality of biological/chemical databases.

48. A system according to Claim 47 further comprising:

means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases including the at least one confidence level that is
5 assigned.

49. A system for integrating a new biological/chemical database with a plurality of biological/chemical databases, comprising:
an entity-relationship model of the plurality of biological/chemical databases
10 that links at least some related entities in at least two of the biological/chemical databases;
an entity-relationship model of the new biological/chemical database;
means for identifying related entities in the entity-relationship model of the new biological/chemical database and the entity-relationship model of the plurality of
15 biological/chemical databases; and
means for linking at least two of the related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

20 50. A system according to Claim 49 wherein the entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in the at least two of the biological/chemical databases provides an ontology network and wherein the entity-relationship model for the new biological/chemical database represents an ontology.

25 51. A system according to Claim 49 wherein the related entities are identical entities and wherein the means for linking comprises means for merging the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases and
30 the new biological/chemical database.

52. A system according to Claim 51 wherein the means for merging further comprises means for establishing a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical

databases and the new biological/chemical database, a respective alias of which refers to a respective one of the at last two of the identical entities that are identified.

53. A system according to Claim 52 wherein the new biological/chemical database is an updated version of one of the plurality of biological/chemical databases, the system further comprising:

means for identifying at least one entity in the one of the plurality of biological/chemical databases that has been deleted from the updated version of the one of the plurality of biological/chemical databases; and

10 means for removing an alias that is associated with the at least one entity that has been removed.

54. A system according to Claim 53 further comprising:

means for splitting at least one entity in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database based upon the alias that was removed.

55. A system according to Claim 49 further comprising:

means for identifying entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database; and

means for adding at least one new entity to the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database that corresponds to the entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

56. A system according to Claim 49 further comprising:

means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response to a query to thereby obtain query results that are based on the entity-

relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

57. A system according to Claim 49 further comprising:

5 means for traversing the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response to a query and in response to at least one path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

10

58. A system according to Claim 56 further comprising:

means for storing the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database as at least one new relationship in the entity-relationship
15 model that integrates the plurality of biological/chemical databases and the new biological/chemical database to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

20 59. A system according to Claim 49 further comprising:

means for maintaining an image of the entity-relationship model of the plurality of biological/chemical databases before the at least two of the related entities are linked.

25 60. A system according to Claim 54 further comprising:

means for comparing the image of the entity-relationship model of the plurality of biological/chemical databases before the at least two of the related entities are linked and the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

30

61. A system according to Claim 49 wherein the entity-relationship model of the new biological/chemical database does not include relationships therein.

62. A system for querying a plurality of biological/chemical databases, each of which includes records for a plurality of biological/chemical entities, the system comprising:

an integrated entity-relationship model of the plurality of biological/chemical
5 databases that links at least some related entities in at least two of the biological/chemical databases; and

means for traversing the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of
10 biological/chemical databases.

63. A system according to Claim 62 wherein the means for traversing comprises:

means for traversing the integrated entity-relationship model of the plurality of
15 biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending entity to thereby identify relationships between the starting entity and the ending entity that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

20 64. A system according to Claim 62 wherein the means for traversing comprises:

means for traversing the integrated entity-relationship model of the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity to thereby identify relationships
25 between the starting entity and the plurality of ending entities that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

65. A system according to Claim 62 wherein the means for traversing comprises:

30 means for traversing the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query and in response to at least one path rule to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

66. A system according to Claim 65 wherein the at least one path rule specifies a type of path to use in traversing through the plurality of entities, a type of path not to use in traversing through the plurality of entities, a type of ending entity that can be included in the query results, a type of ending entity that is not to be included in the query results, a type or class of relationship that is to be used in traversing through the plurality of entities, a type or class of relationship not to be used in traversing through the plurality of entities and/or a confidence level to be achieved in traversing through the plurality of entities.

67. A system according to Claim 65 further comprising storing the query and the path rule for reuse.

68. A system according to Claim 62 further comprising:
means for storing the query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases as at least one new relationship in the integrated entity-relationship model of the plurality of biological/chemical databases to thereby store knowledge that was derived from the query in the integrated entity-relationship model of the plurality of biological/chemical databases.

69. A system according to Claim 62 further comprising:
means for assigning a confidence level to at least one of the relationships in the integrated entity-relationship model of the plurality of biological/chemical databases.

70. A system according to Claim 69 further comprising:
means for traversing the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases including the at least one confidence level that is assigned.

71. A computer program product that is configured to integrate a plurality of biological/chemical databases, the computer program product comprising a

computer usable storage medium having computer-readable program code embodied in the medium, the computer-readable program code comprising:

computer-readable program code that is configured to obtain an entity-relationship model for each of the plurality of biological/chemical databases;

5 computer-readable program code that is configured to identify related entities in the entity-relationship models of at least two of the biological/chemical databases; and

computer-readable program code that is configured to link at least two of the related entities that are identified, to thereby create an entity-relationship model that
10 integrates the plurality of biological/chemical databases.

72. A computer program product according to Claim 71 wherein at least one of the plurality of databases represents an ontology and wherein the entity-relationship model that integrates the plurality of biological/chemical databases
15 creates an ontology network.

73. A computer program product according to Claim 71 wherein the related entities are identical entities and wherein the computer-readable program code that is configured to link comprises computer-readable program code that is
20 configured to merge the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of biological/chemical databases.

74. A computer program product according to Claim 73 wherein the
25 computer-readable program code that is configured to merge further comprises computer-readable program code that is configured to establish a plurality of aliases for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases, a respective alias of which refers to a respective one of the at least two of the identical entities that are identified.

30

75. A computer program product according to Claim 71 further comprising:

computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases in

response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

76. A computer program product according to Claim 75 wherein the
5 computer-readable program code that is configured to traverse comprises:
computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting entity and the ending entity to thereby identify relationships between the starting
10 entity and the ending entity that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

77. A computer program product according to Claim 75 wherein the
computer-readable program code that is configured to traverse comprises:
15 computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the starting entity to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the entity-relationship model that
20 integrates the plurality of biological/chemical databases.

78. A computer program product according to Claim 75 wherein the
computer-readable program code that is configured to traverse comprises:
computer-readable program code that is configured to traverse the entity-
25 relationship model that integrates the plurality of biological/chemical databases in response to a query and in response to at least one path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases.

79. A computer program product according to Claim 78 wherein the at
30 least one path rule specifies a type of path to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of path not to use in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type of ending entity that

can be included in the query results, a type of ending entity that is not to be included in the query results, a type or class of relationship to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases, a type or class of relationship that is not to be used in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases and/or a confidence level to be achieved in traversing through the entity-relationship model that integrates the plurality of biological/chemical databases.

80. A computer program product according to Claim 78 further comprising computer-readable program code that is configured to store the query and the path rule for reuse.

81. A computer program product according to Claim 75 further comprising:
computer-readable program code that is configured to store the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases as at least one new relationship in the entity-relationship model that integrates the plurality of biological/chemical databases to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of biological/chemical databases.

82. A computer program product according to Claim 75 further comprising:
computer-readable program code that is configured to assign a confidence level to at least one of the relationships in the entity-relationship model that integrates the plurality of biological/chemical databases.

83. A computer program product according to Claim 82 further comprising:
computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases including the at least one confidence level that is assigned.

84. A computer program product that is configured to integrate a new biological/chemical database with a plurality of biological/chemical databases, the computer program product comprising a computer usable storage medium having
5 computer-readable program code embodied in the medium, the computer-readable program code comprising:

- an entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in at least two of the biological/chemical databases;
- 10 an entity-relationship model of the new biological/chemical database; computer-readable program code that is configured to identify related entities in the entity-relationship model of the new biological/chemical database and the entity-relationship model of the plurality of biological/chemical databases; and
- computer-readable program code that is configured to link at least two of the
15 related entities that are identified, to thereby create an entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

85. A computer program product according to Claim 84 wherein the
20 entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in the at least two of the biological/chemical databases provides an ontology network and wherein the entity-relationship model for the new biological/chemical database represents an ontology.

86. A computer program product according to Claim 84 wherein the
25 related entities are identical entities and wherein the computer-readable program code that is configured to link comprises computer-readable program code that is configured to merge the at least two of the identical entities that are identified into a single entity in the entity-relationship model that integrates the plurality of
30 biological/chemical databases and the new biological/chemical database.

87. A computer program product according to Claim 86 wherein the computer-readable program code that is configured to merge further comprises computer-readable program code that is configured to establish a plurality of aliases

for the single entity in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database, a respective alias of which refers to a respective one of the at last two of the identical entities that are identified.

5

88. A computer program product according to Claim 87 wherein the new biological/chemical database is an updated version of one of the plurality of biological/chemical databases, the computer program product further comprising:

computer-readable program code that is configured to identify at least one
10 entity in the one of the plurality of biological/chemical databases that has been deleted from the updated version of the one of the plurality of biological/chemical databases; and

computer-readable program code that is configured to remove an alias that is associated with the at least one entity that has been removed.

15

89. A computer program product according to Claim 88 further comprising:

computer-readable program code that is configured to split at least one entity
in the entity-relationship model that integrates the plurality of biological/chemical
20 databases and the new biological/chemical database based upon the alias that was removed.

90. A computer program product according to Claim 84 further comprising:

25 computer-readable program code that is configured to identify entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database; and

computer-readable program code that is configured to add at least one new
30 entity to the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database that corresponds to the entities in the new biological/chemical database that do not correspond to at least one of the entities in the entity-relationship model that

integrates the plurality of biological/chemical databases and the new biological/chemical database.

91. A computer program product according to Claim 84 further comprising:

computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response to a query to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

92. A computer program product according to Claim 84 further comprising:

computer-readable program code that is configured to traverse the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database in response to a query and in response to at least one path rule to thereby obtain query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

20

93. A computer program product according to Claim 91 further comprising:

computer-readable program code that is configured to store the query results that are based on the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database as at least one new relationship in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database to thereby store knowledge that was derived from the query in the entity-relationship model that integrates the plurality of biological/chemical databases and the new biological/chemical database.

94. A computer program products according to Claim 84 further comprising:

computer-readable program code that is configured to maintain an image of the entity-relationship model of the plurality of biological/chemical databases before the at least two of the related entities are linked.

5 95. A computer program product according to Claim 94 further comprising:

 computer-readable program code that is configured to compare the image of the entity-relationship model of the plurality of biological/chemical databases before the at least two of the related entities are linked and the entity relationship mode that
10 integrates the plurality of biological chemical databases and the new biological/chemical database.

 96. A computer program product according to Claim 84 wherein the entity-relationship model of the new biological/chemical database does not include
15 relationships therein.

 97. A computer program product that is configured to query a plurality of biological/chemical databases, each of which includes records for a plurality of biological/chemical entities, the computer program product comprising a computer
20 usable storage medium having computer-readable program code embodied in the medium, the computer-readable program code comprising:

 an integrated entity-relationship model of the plurality of biological/chemical databases that links at least some related entities in at least two of the biological/chemical databases; and

25 computer-readable program code that is configured to traverse the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

30 98. A computer program product according to Claim 97 wherein the computer-readable program code that is configured to traverse comprises:

 computer-readable program code that is configured to traverse the integrated entity-relationship model of the plurality of biological/chemical databases from a starting entity to an ending entity in response to a query that specifies the starting

entity and the ending entity to thereby identify relationships between the starting entity and the ending entity that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

- 5 99. A computer program product according to Claim 97 wherein the computer-readable program code that is configured to traverse comprises:
 computer-readable program code that is configured to traverse the integrated entity-relationship model of the plurality of biological/chemical databases from a starting entity to a plurality of ending entities in response to a query that specifies the
10 starting entity to thereby identify relationships between the starting entity and the plurality of ending entities that are based on the integrated entity-relationship model of the plurality of biological/chemical databases.

100. A computer program product according to Claim 97 wherein the
15 computer-readable program code that is configured to traverse comprises:
 computer-readable program code that is configured to traverse the integrated entity-relationship model of the plurality of biological/chemical databases in response to a query and in response to at least one path rule to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of
20 biological/chemical databases.

101. A computer program product according to Claim 100 wherein the at least one path rule specifies a type of path to use in traversing through the plurality of entities, a type of path not to use in traversing through the plurality of entities, a type
25 of ending entity that can be included in the query results, a type of ending entity that is not to be included in the query results, a type or class of relationship that is to be used in traversing through the plurality of entities, a type or class of relationship not to be used in traversing through the plurality of entities and/or a confidence level to be achieved in traversing through the plurality of entities.

30

102. A computer program products according to Claim 100 further comprising computer-readable program code that is configured to store the query and the path rule for reuse.

103. A computer program product according to Claim 97 further comprising:

computer-readable program code that is configured to store the query results that are based on the integrated entity-relationship model of the plurality of
5 biological/chemical databases as at least one new relationship in the integrated entity-relationship model of the plurality of biological/chemical databases to thereby store knowledge that was derived from the query in the integrated entity-relationship model of the plurality of biological/chemical databases.

10 104. A computer program product according to Claim 97 further comprising:

computer-readable program code that is configured to assign a confidence level to at least one of the relationships in the integrated entity-relationship model of the plurality of biological/chemical databases.

15

105. A computer program product according to Claim 104 further comprising:

computer-readable program code that is configured to traverse the integrated entity-relationship model of the plurality of biological/chemical databases in response
20 to a query to thereby obtain query results that are based on the integrated entity-relationship model of the plurality of biological/chemical databases including the at least one confidence level that is assigned.

106. A bioinformatics data processing system comprising:
25 an ontology network engine that is configured to build an integrated entity-relationship model of a plurality of independent biological/chemical databases, each of which includes records for a plurality of biological/chemical objects, the integrated entity-relationship model comprising:

a plurality of entities, a respective one of which corresponds to a single
30 biological/chemical object, at least some of the entities including a plurality of links, a respective one of which directly or indirectly refers to at least one record in a respective one of the plurality of biological/chemical databases that relates to the single biological/chemical object; and

a plurality of relationships that link the plurality of entities in the entity-relationship model based upon relationships therebetween.

107. A system according to Claim 106 further comprising:
5 a metadata database that is configured to store therein the integrated entity-relationship model of the plurality of independent biological/chemical databases.

108. A system according to Claim 106 further comprising:
a loader that is configured to load an independent entity-relationship model of
10 each of the independent biological/chemical databases into the ontology network engine.

109. A system according to Claim 108 wherein the loader is configured to load an independent entity-relationship model of each of the independent
15 biological/chemical databases into the ontology network engine in a typeless format.

110. A system according to Claim 108 in combination with the plurality of independent biological/chemical databases.

20 111. A system according to Claim 106 further comprising:
a query tool that is configured to traverse the integrated entity-relationship model in response to a query to thereby obtain query results that are based on the integrated entity-relationship model.

25 112. A system according to Claim 111 wherein the query tool is a Web-based query tool.

113. A system according to Claim 106 further comprising:
a virtual experiment tool that is configured to conduct virtual experiments on
30 the integrated entity-relationship model.

114. A system according to Claim 106 further comprising:
a discovery tool that is configured to discover biological/chemical knowledge from the integrated entity-relationship model.

115. A system according to Claim 106 wherein the ontology network engine runs on a plurality of data processing systems that are configured in a peer-to-peer configuration.

5

116. A bioinformatics data structure comprising:

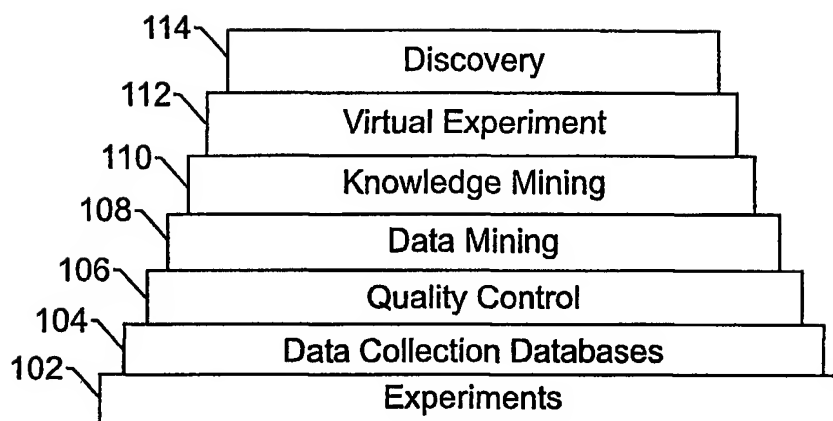
an integrated entity-relationship model of a plurality of independent biological/chemical databases, each of which includes records for a plurality of biological/chemical objects, the integrated entity-relationship model comprising:

- 10 a plurality of entities, a respective entity of which corresponds to a single biological/chemical object, at least some of the entities including a plurality of links, a respective one of which directly or indirectly refers to at least one record in a respective one of the plurality of biological/chemical databases that relates to the single biological/chemical object; and
- 15 a plurality of relationships that link the plurality of entities in the entity-relationship model based upon relationships therebetween.

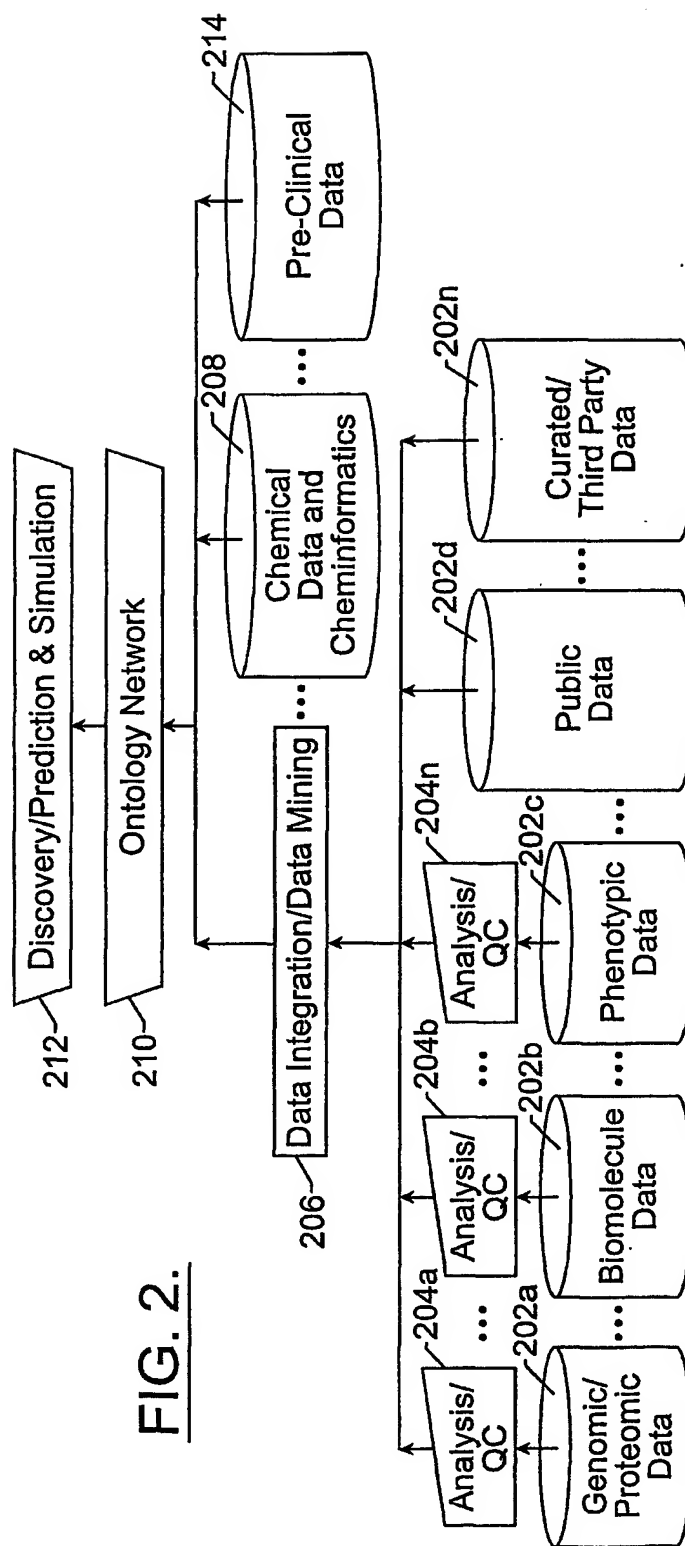
117. A data structure according to Claim 116 further comprising:

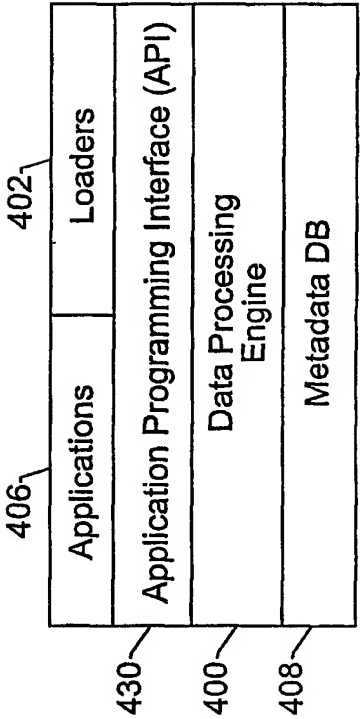
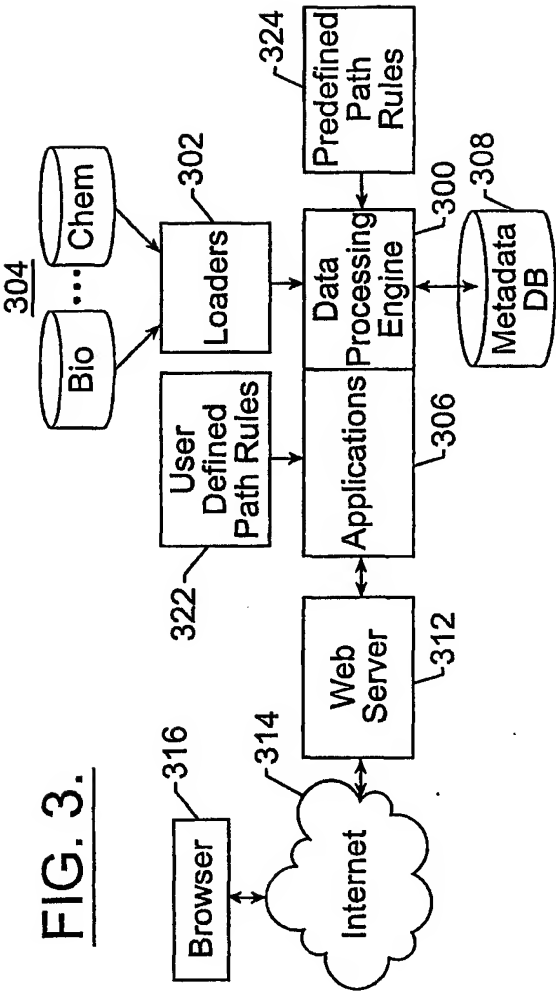
- an independent entity-relationship model of each of the independent
- 20 biological/chemical databases.

1/38

FIG. 1.

2/38





4/38

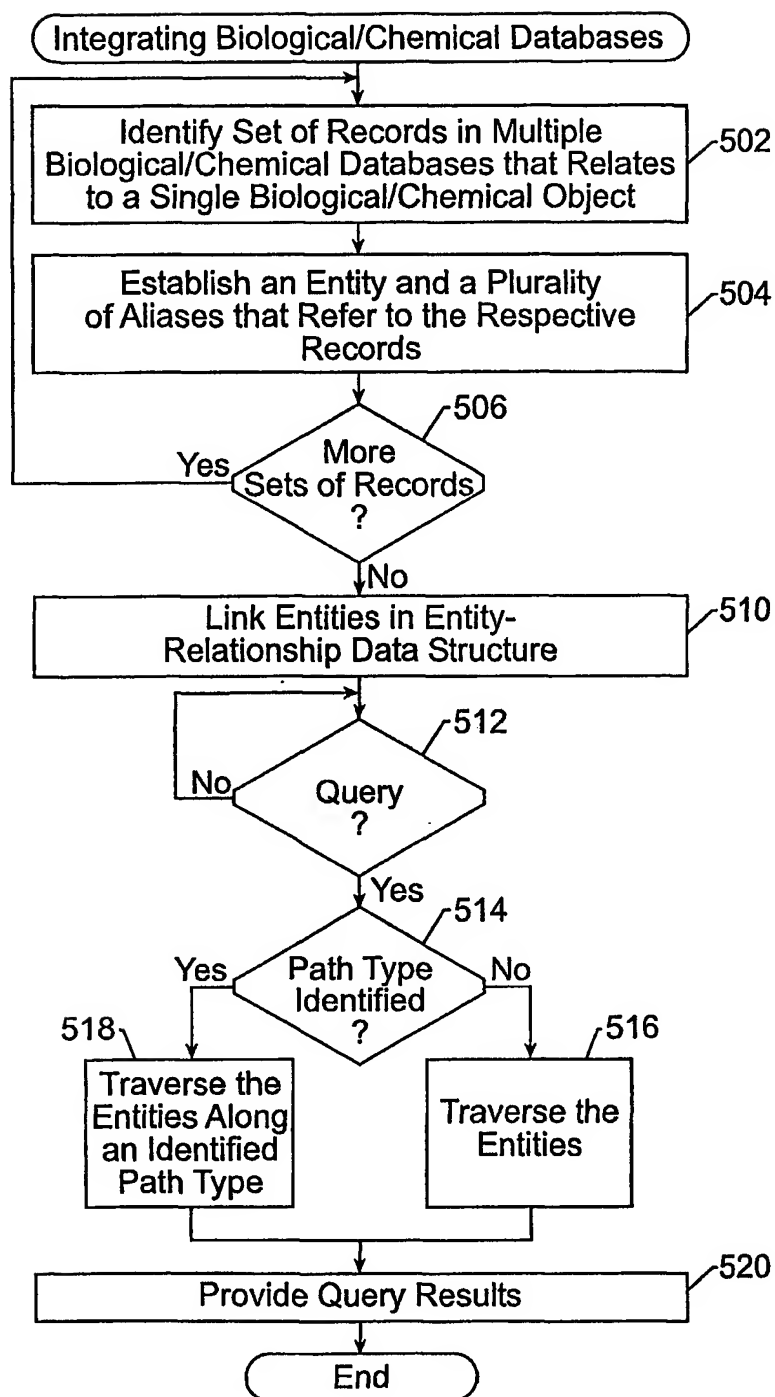
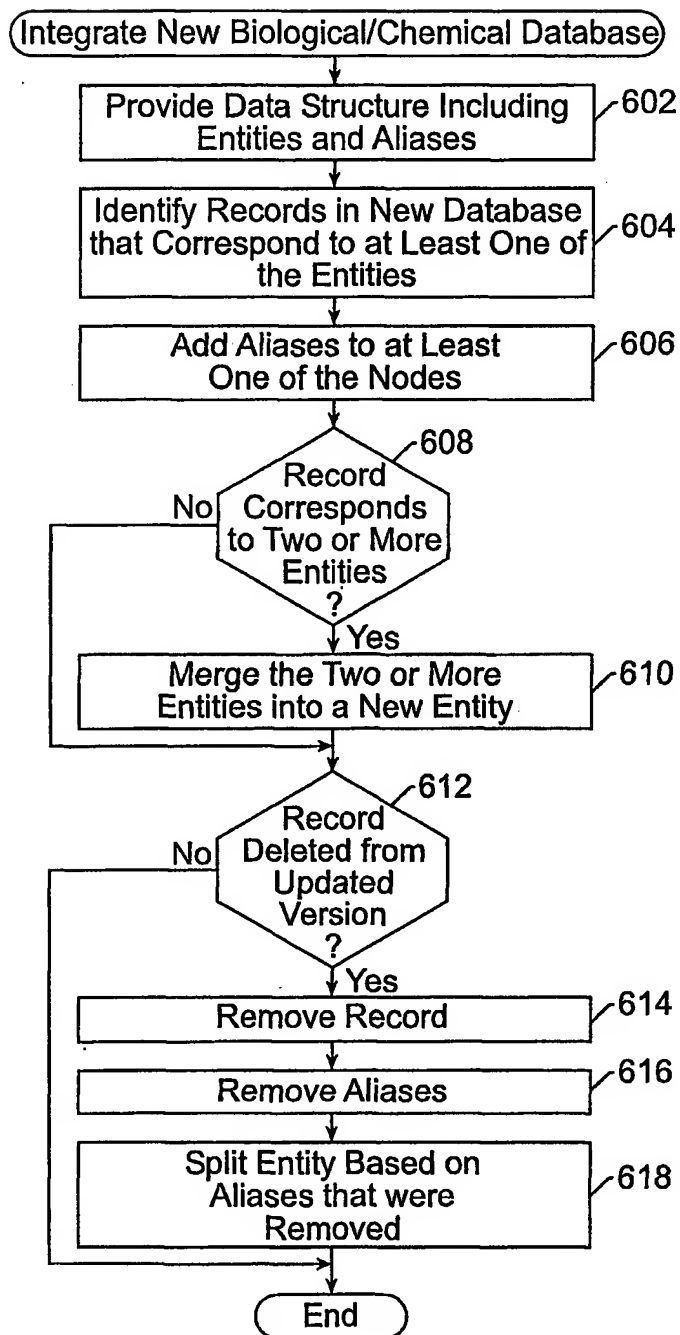
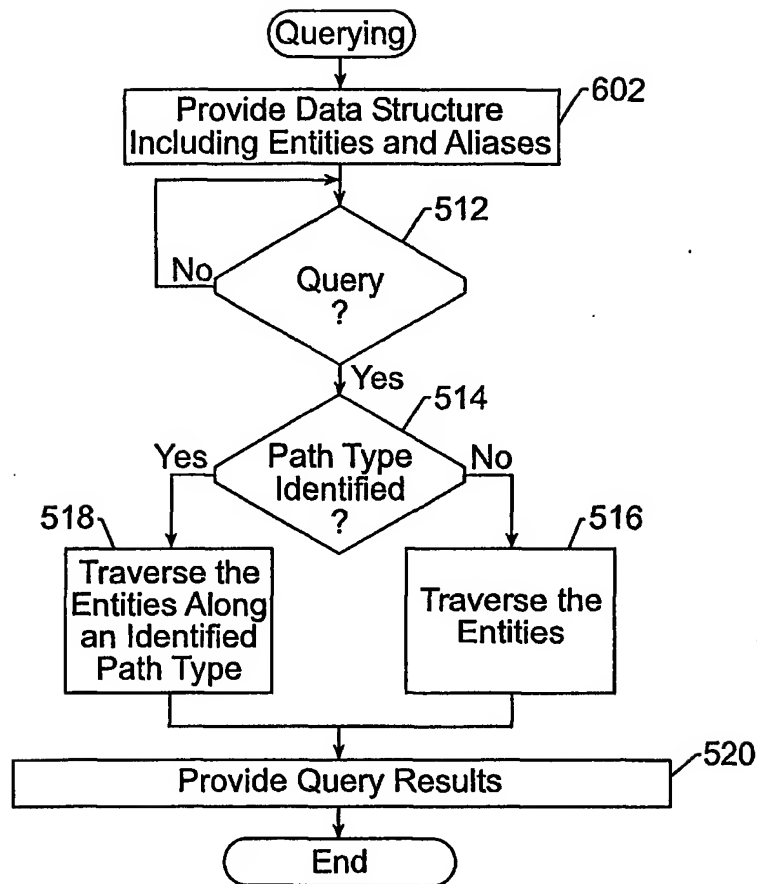


FIG. 5.

5/38

**FIG. 6.**

6/38

FIG. 7.

7/38

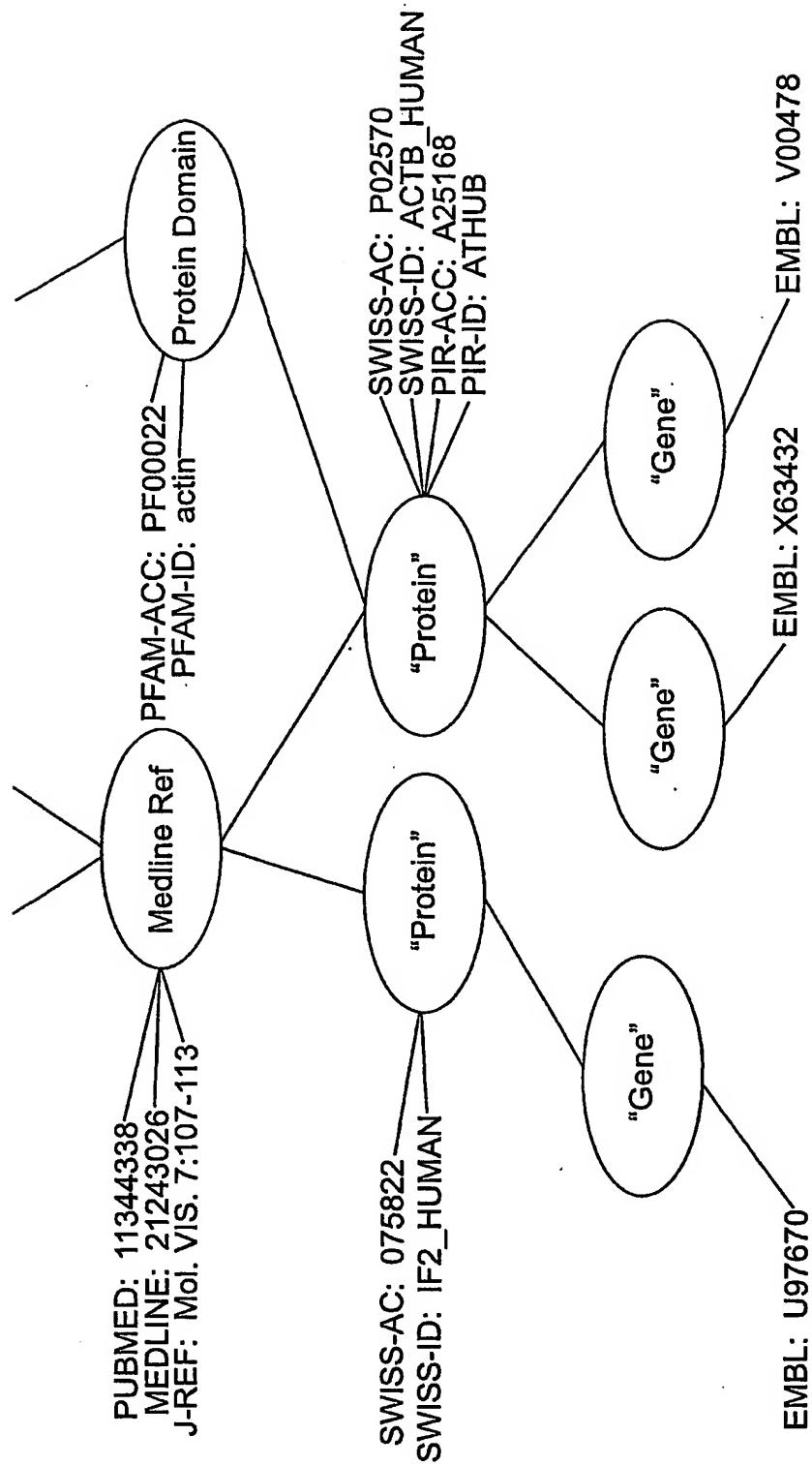
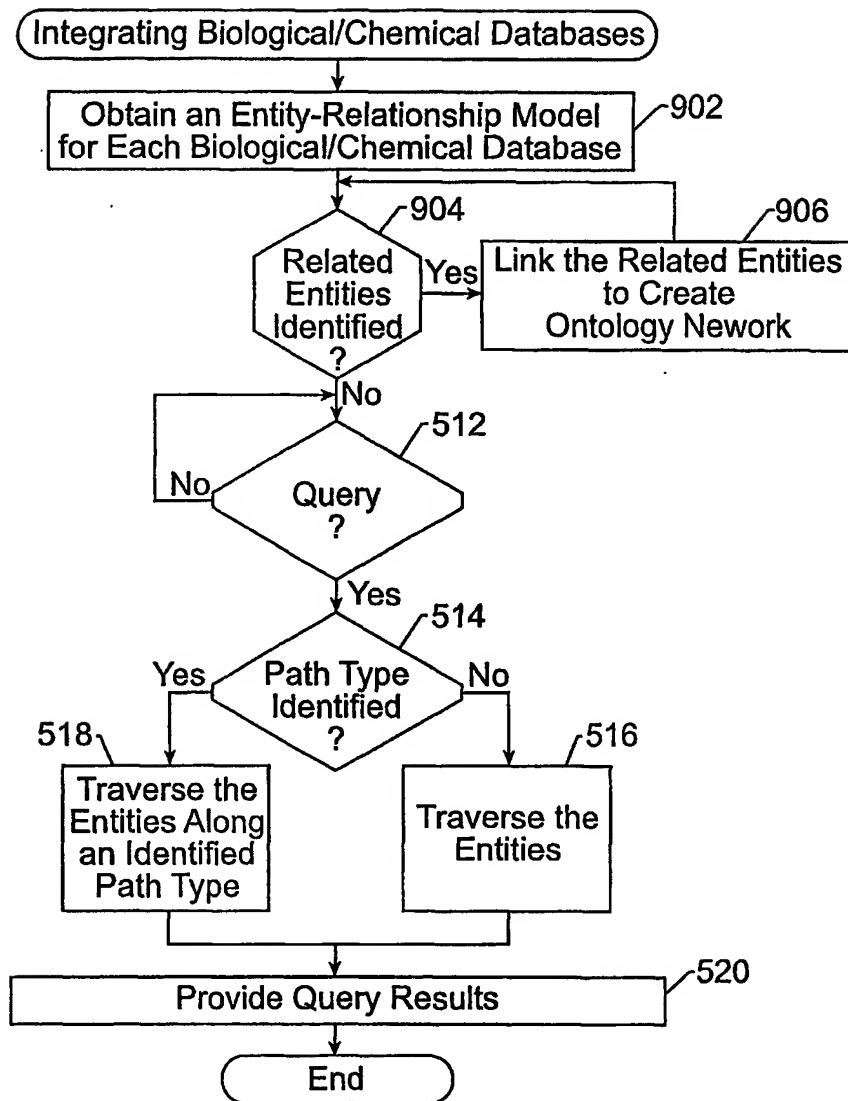
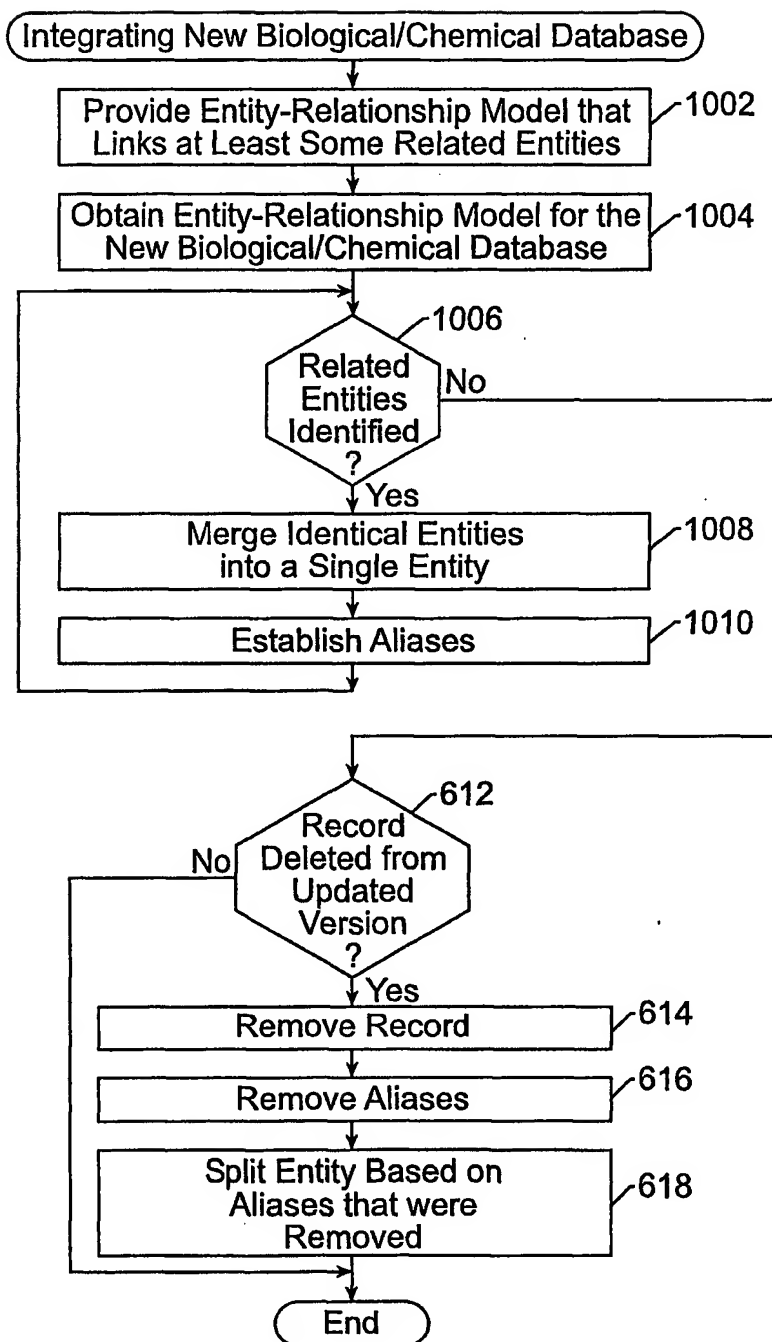


FIG. 8.

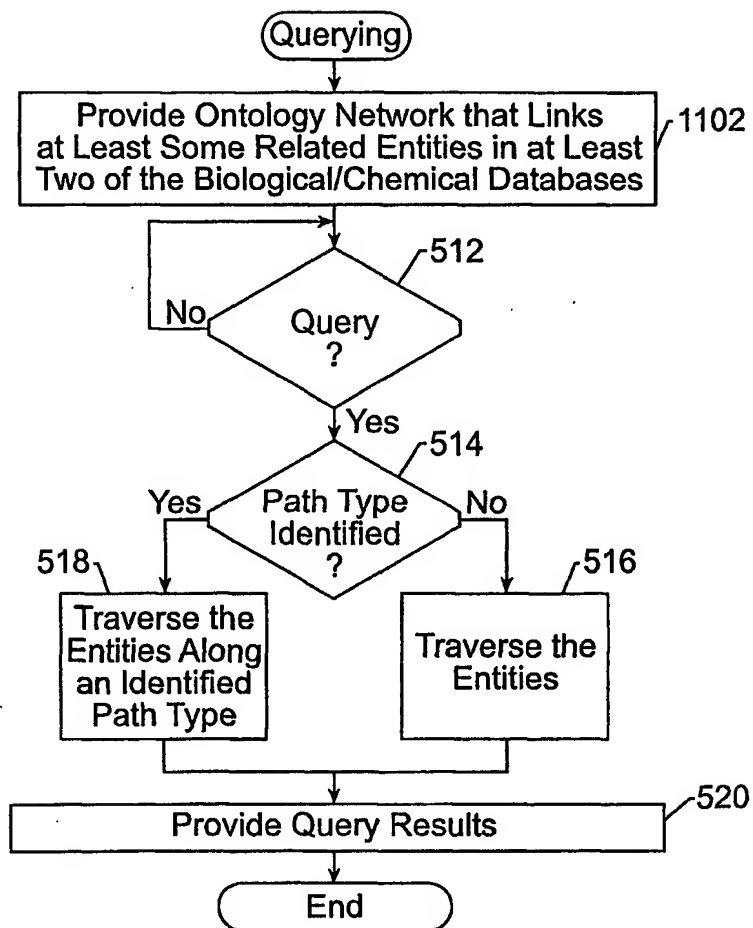
8/38

FIG. 9.

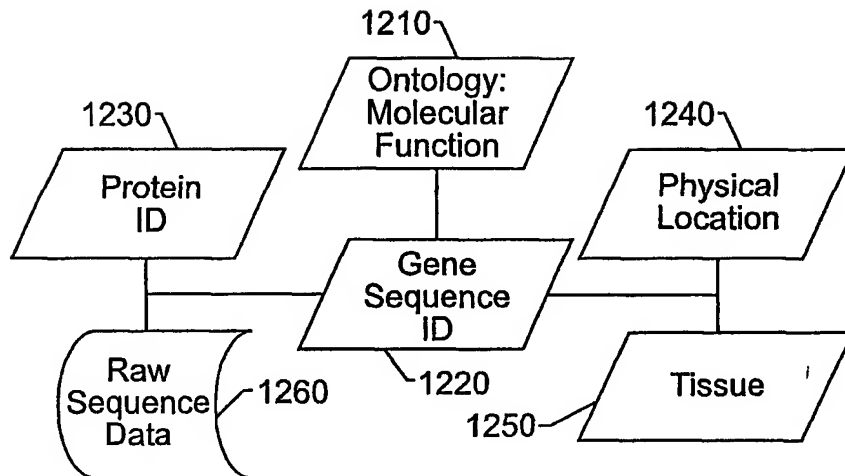
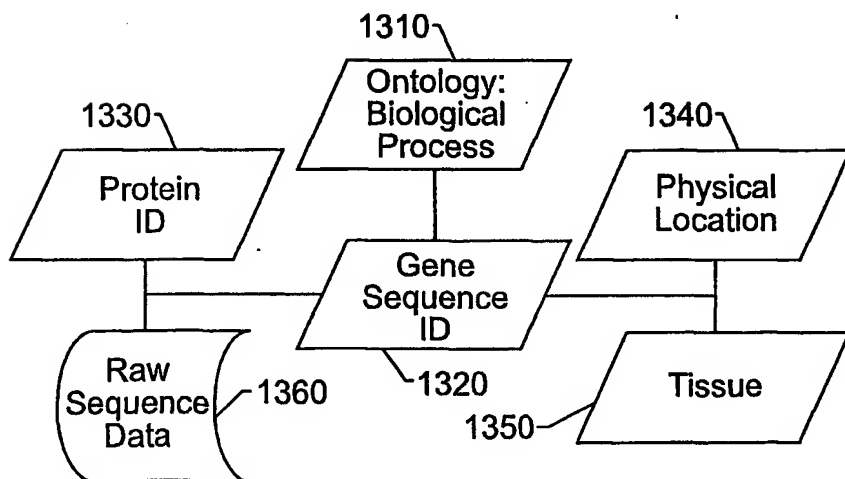
9/38

FIG. 10.

10/38

FIG. 11.

11/38

FIG. 12.FIG. 13.

12/38

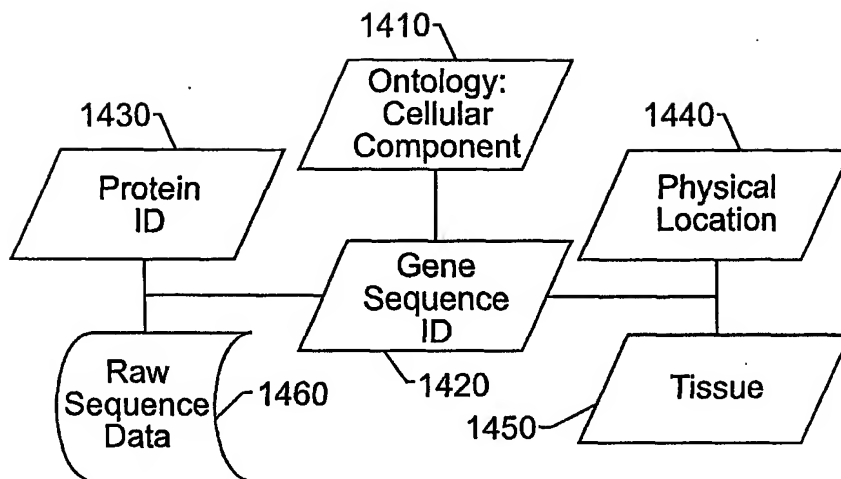


FIG. 14.

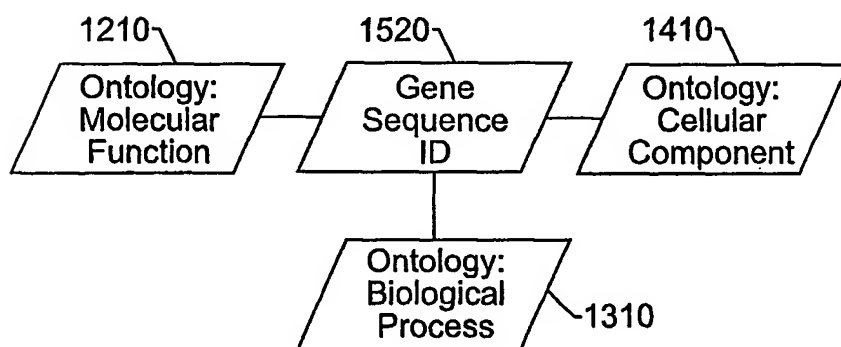
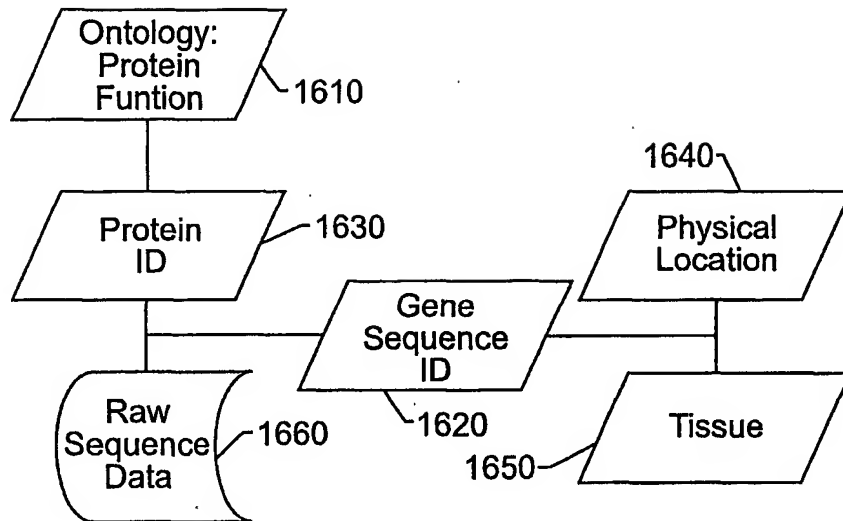
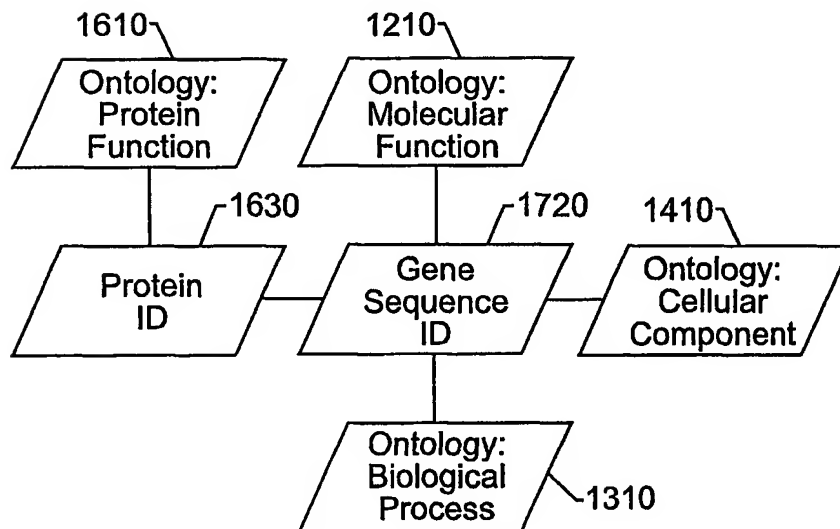
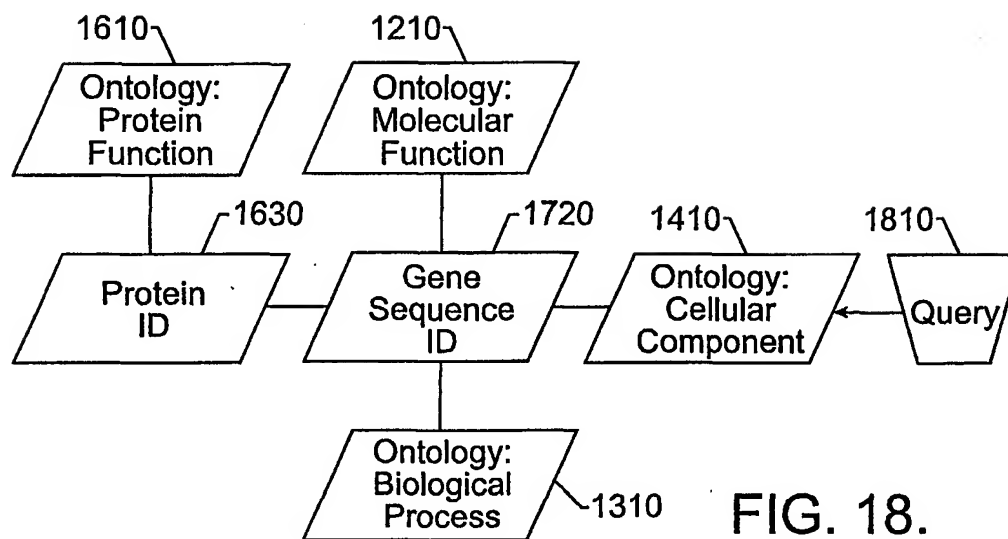
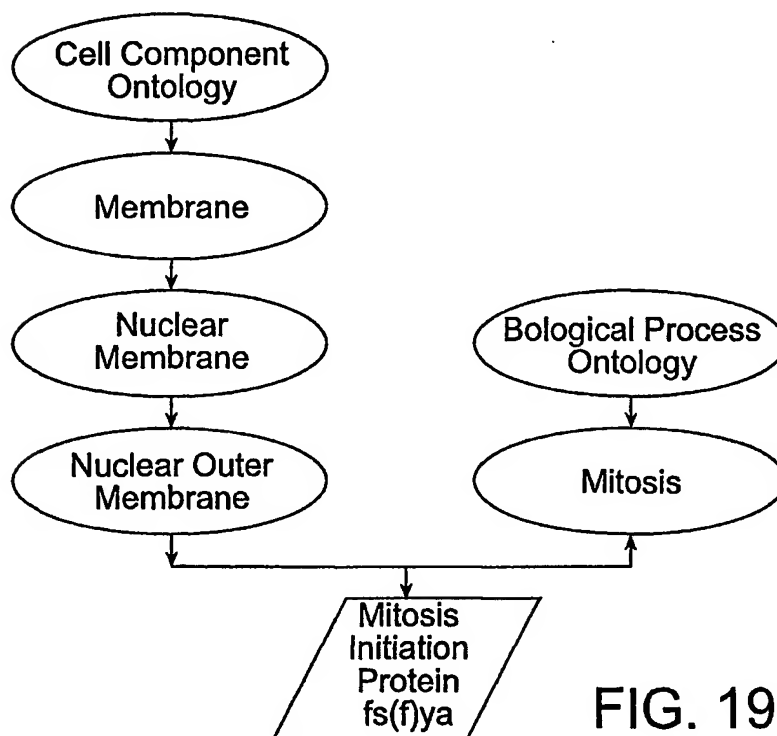


FIG. 15.

13/38

FIG. 16.FIG. 17.

14/38

FIG. 18.FIG. 19.

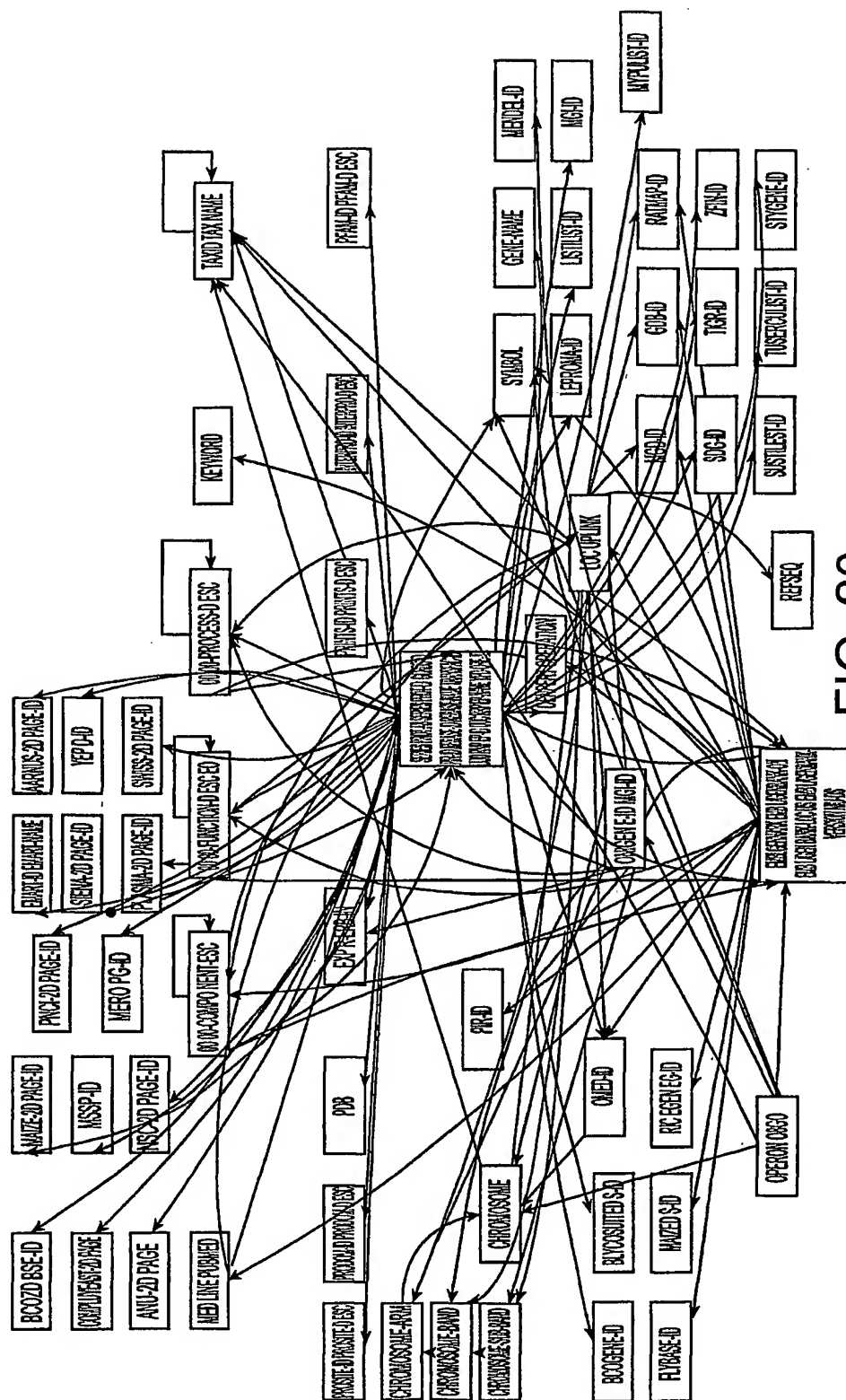
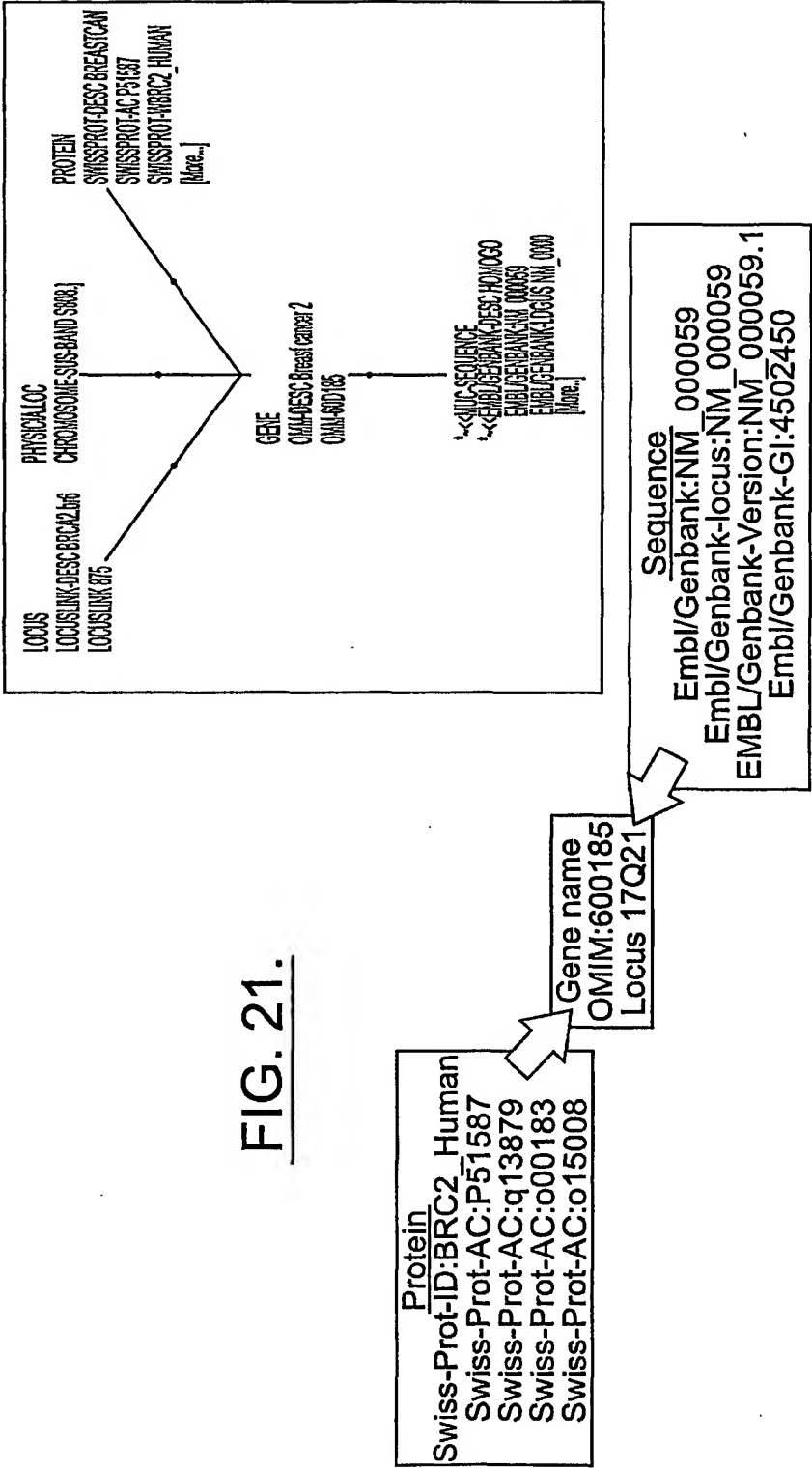


FIG. 20.

FIG. 21.



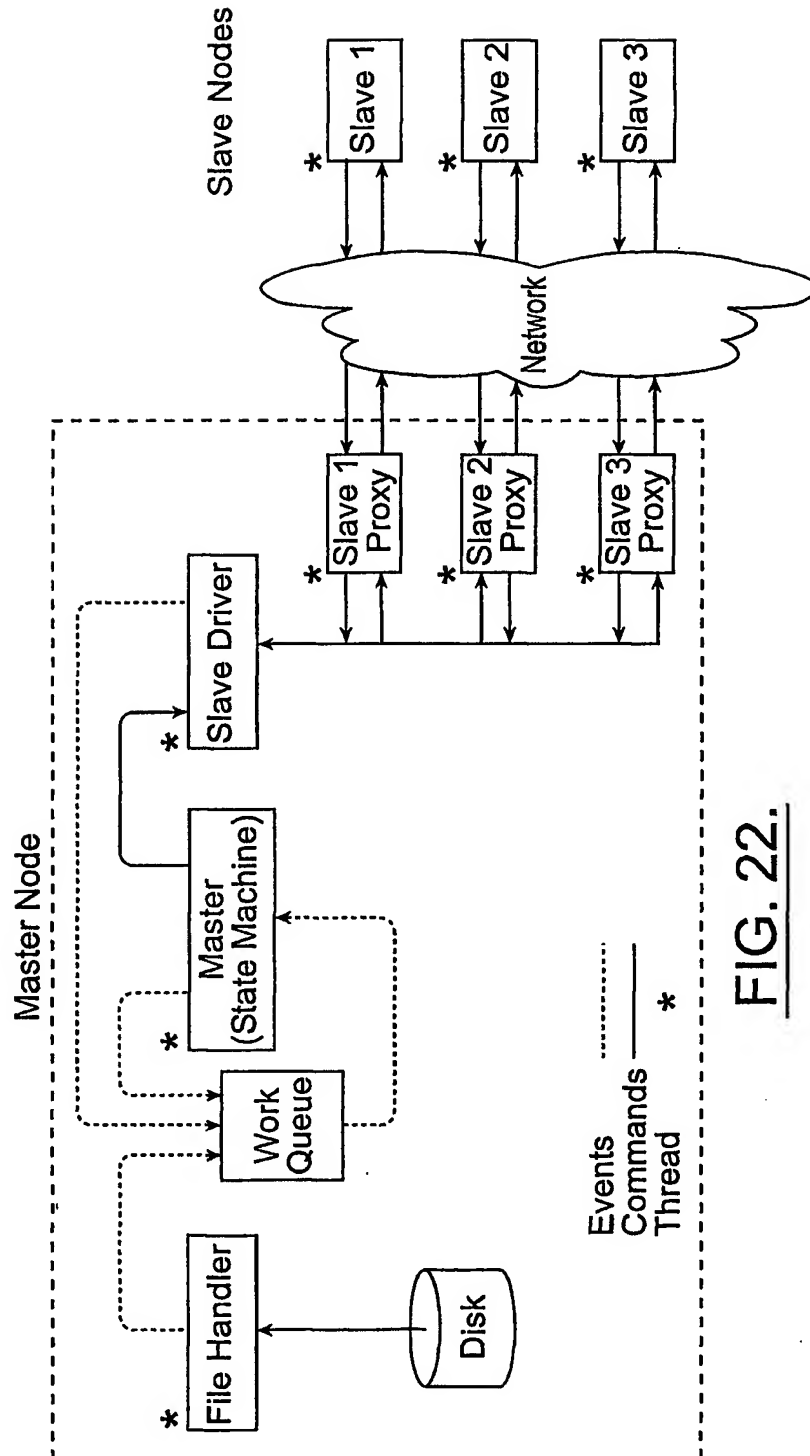


FIG. 22.

18/38

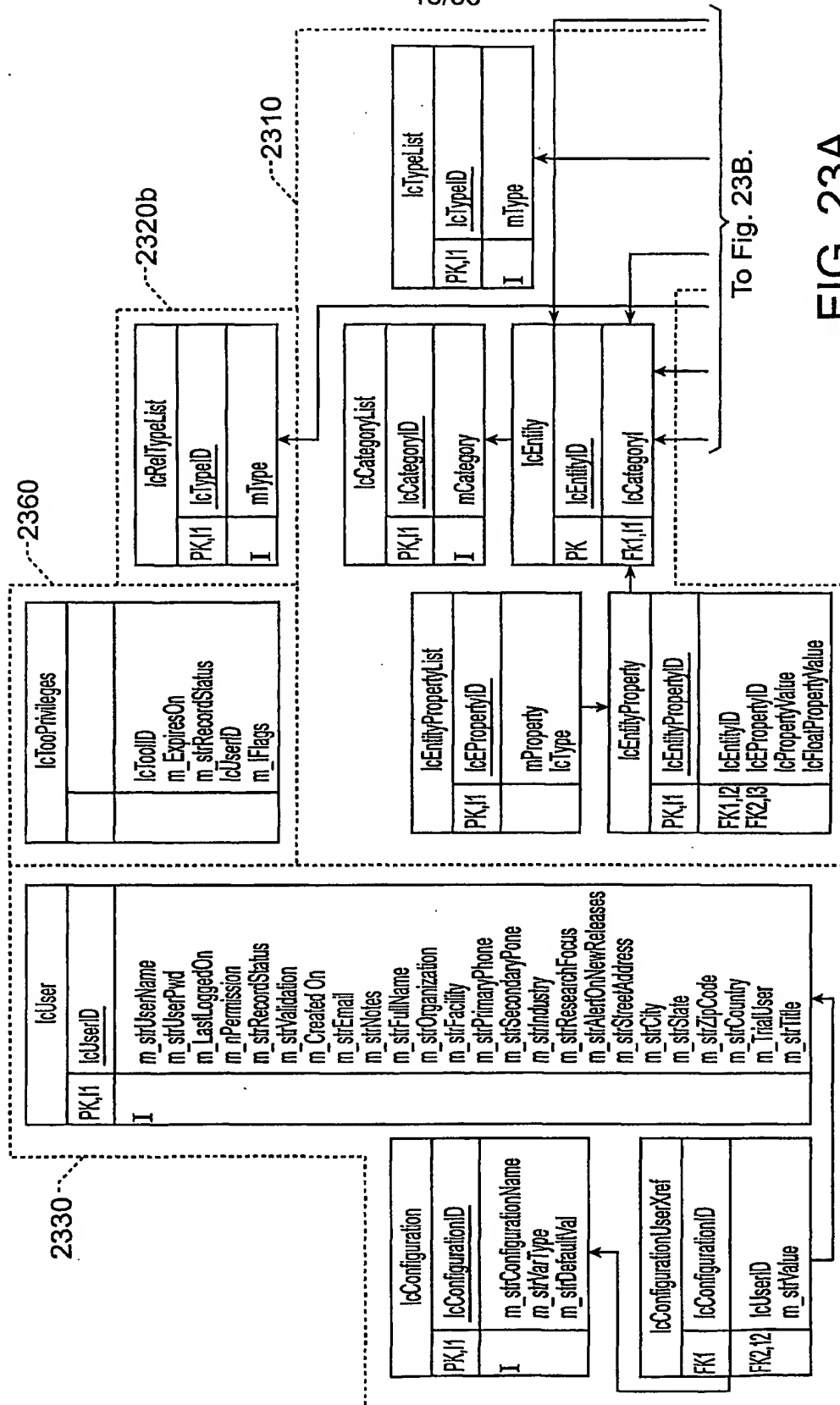
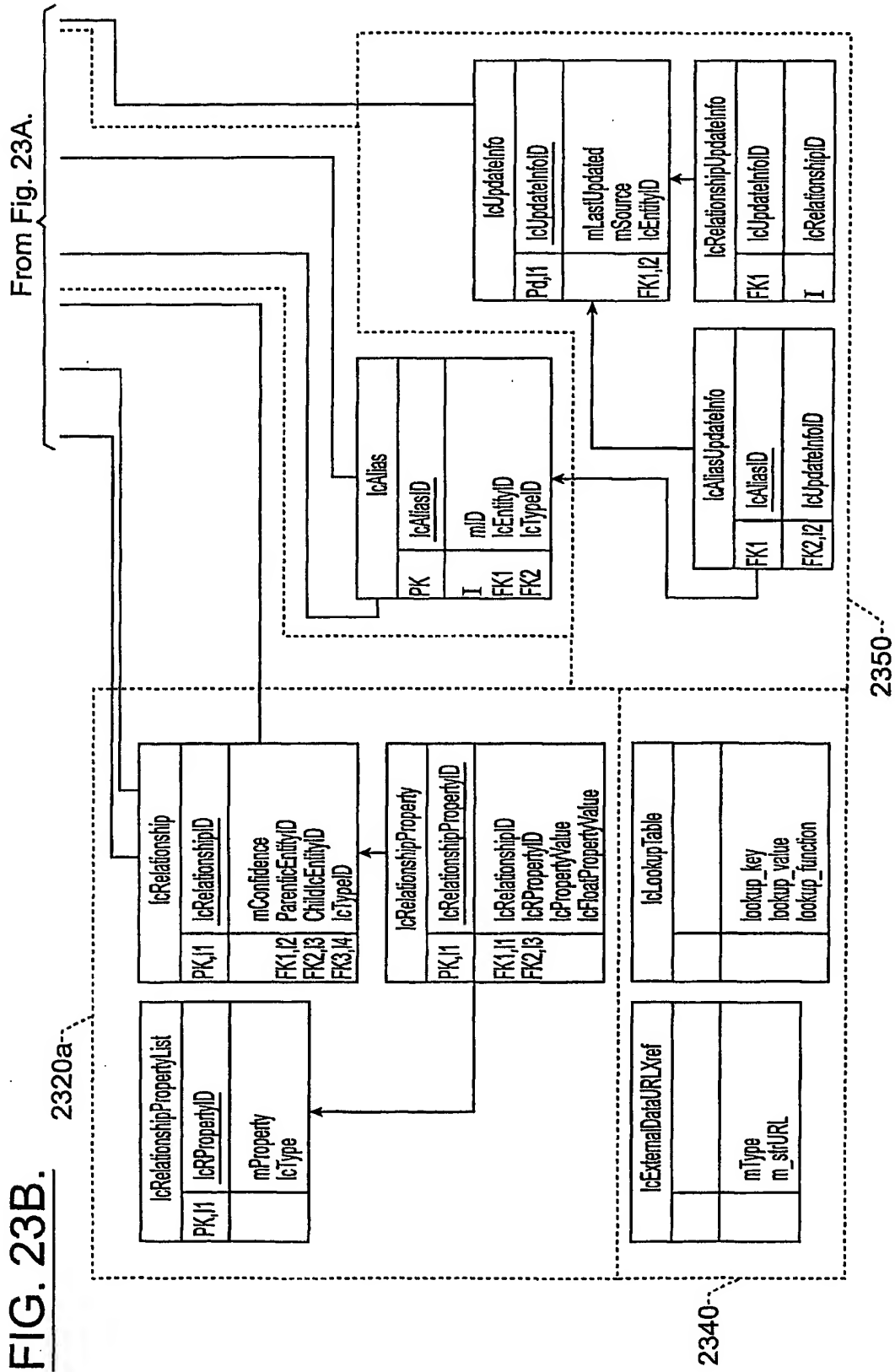


FIG. 23A.

19/38



20/38

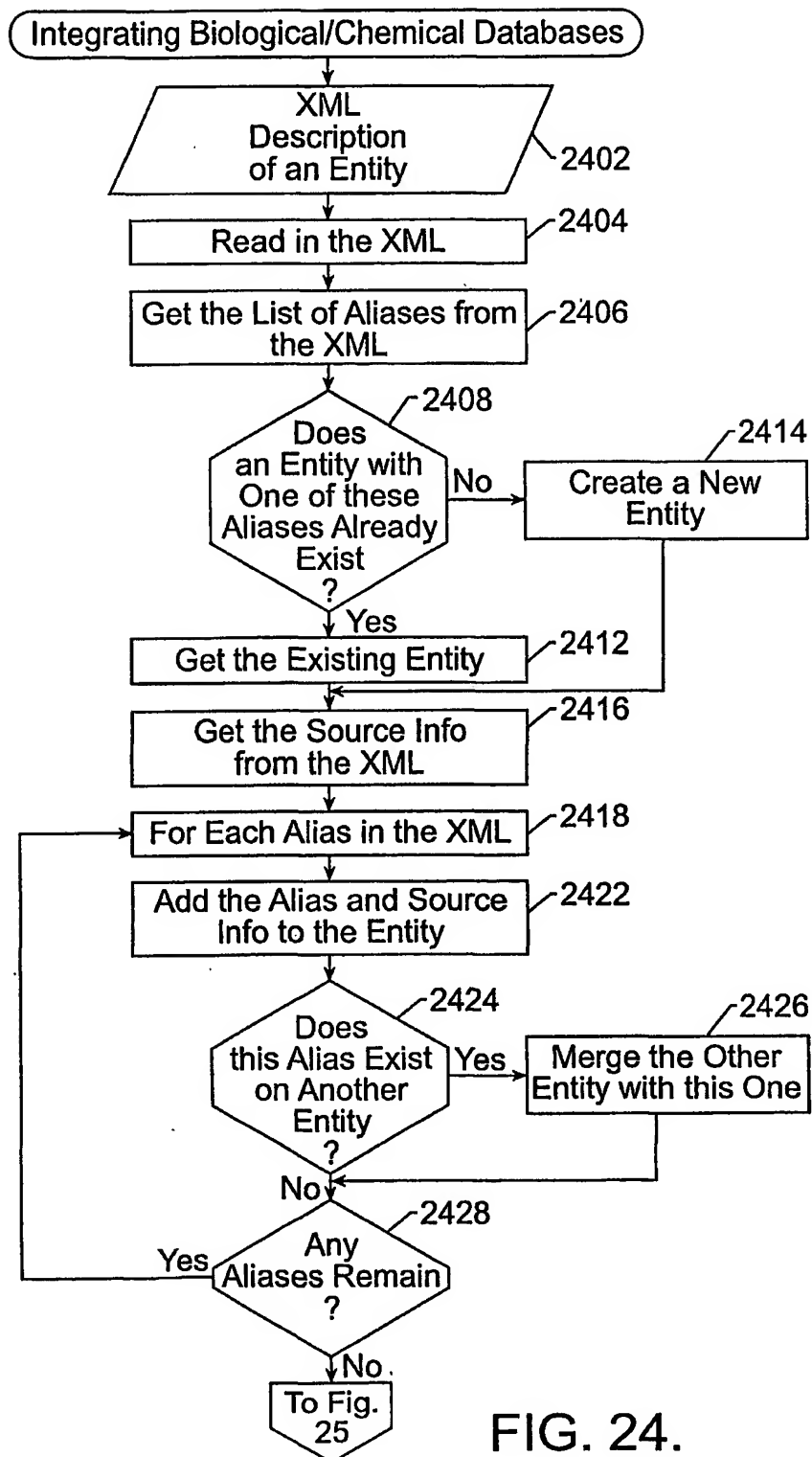


FIG. 24.

21/38

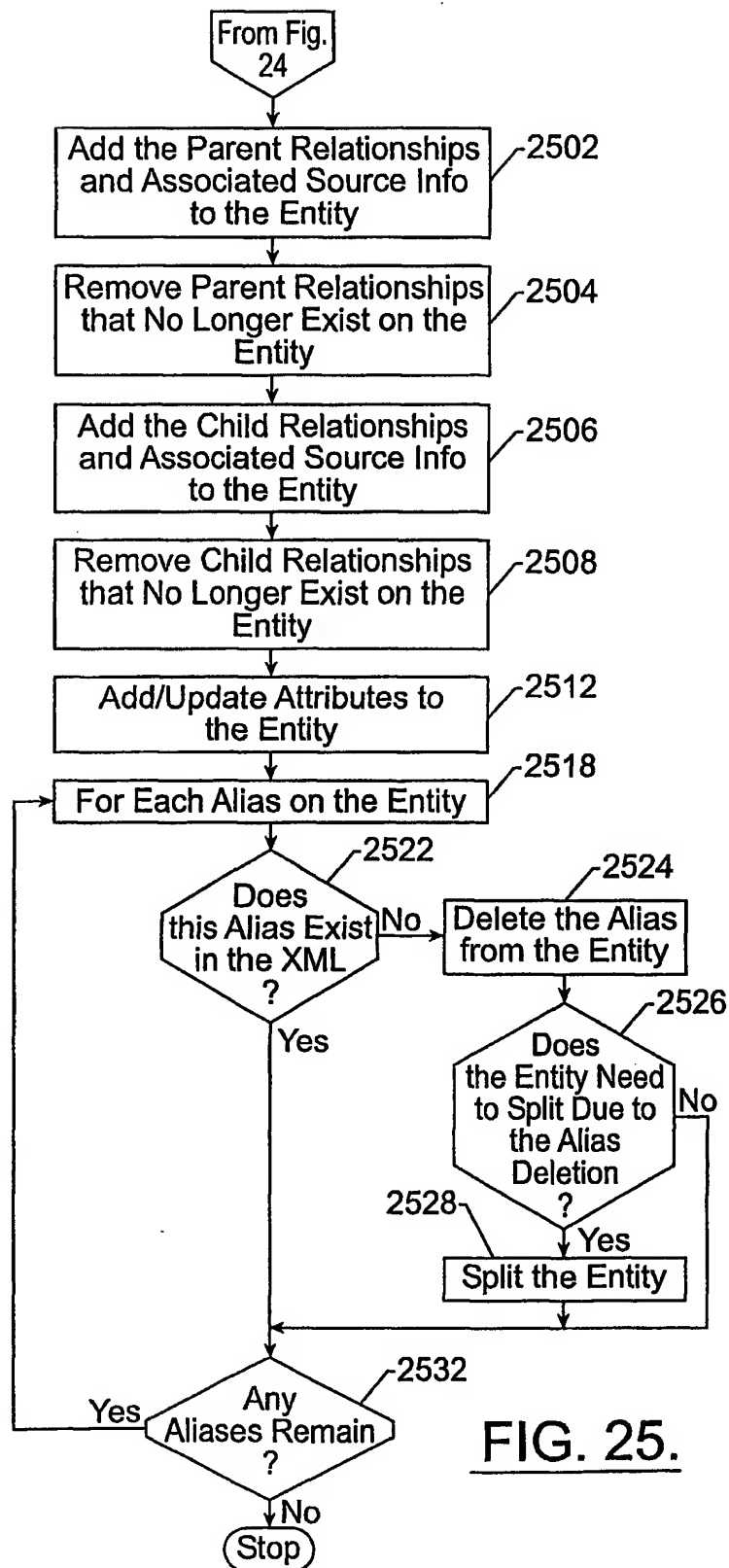
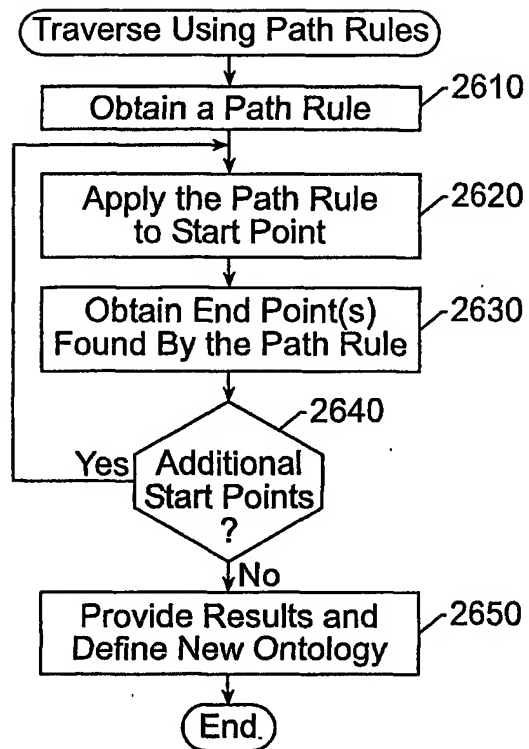
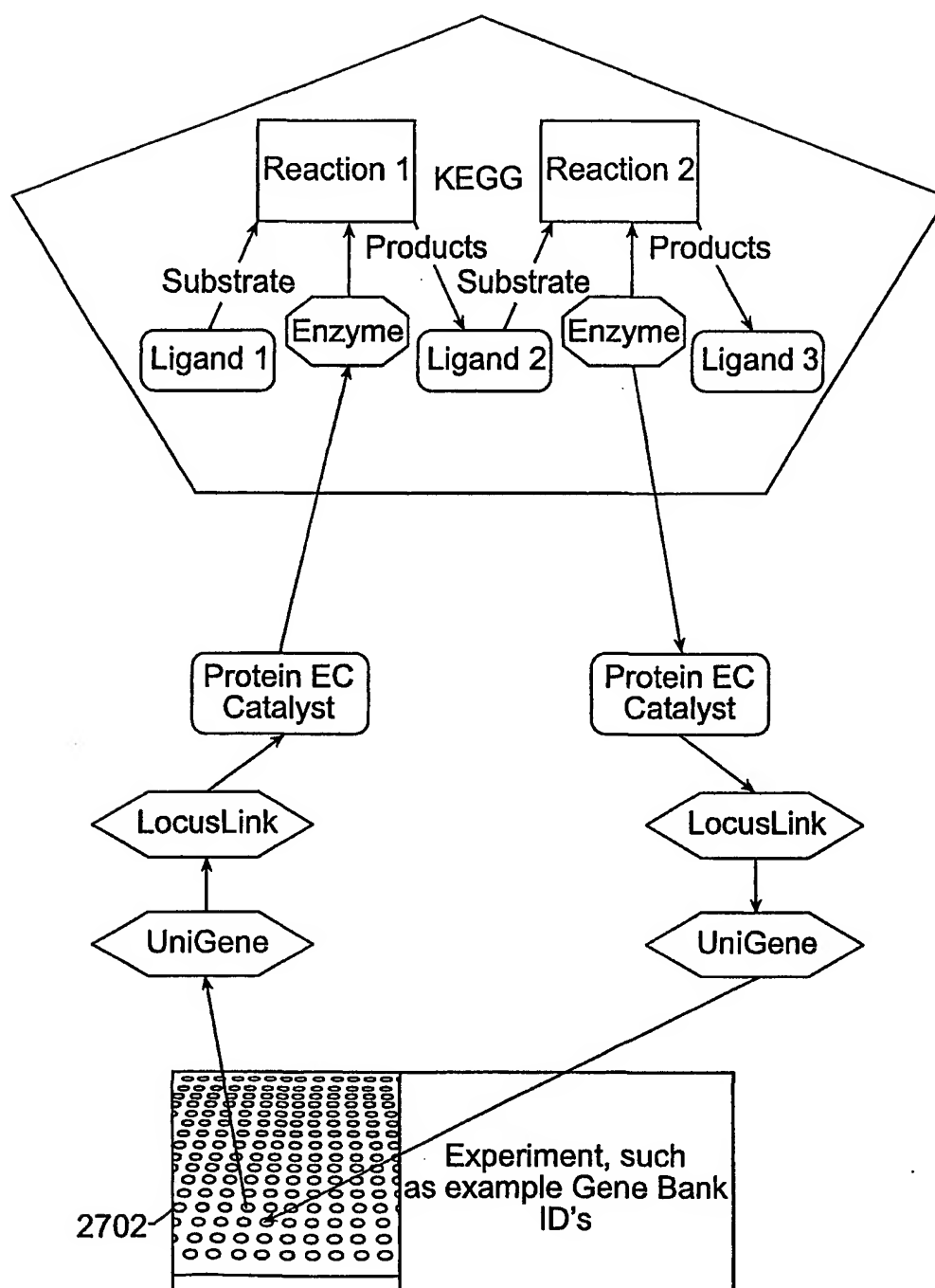


FIG. 25.

22/38

FIG. 26.

23/38

FIG. 27.

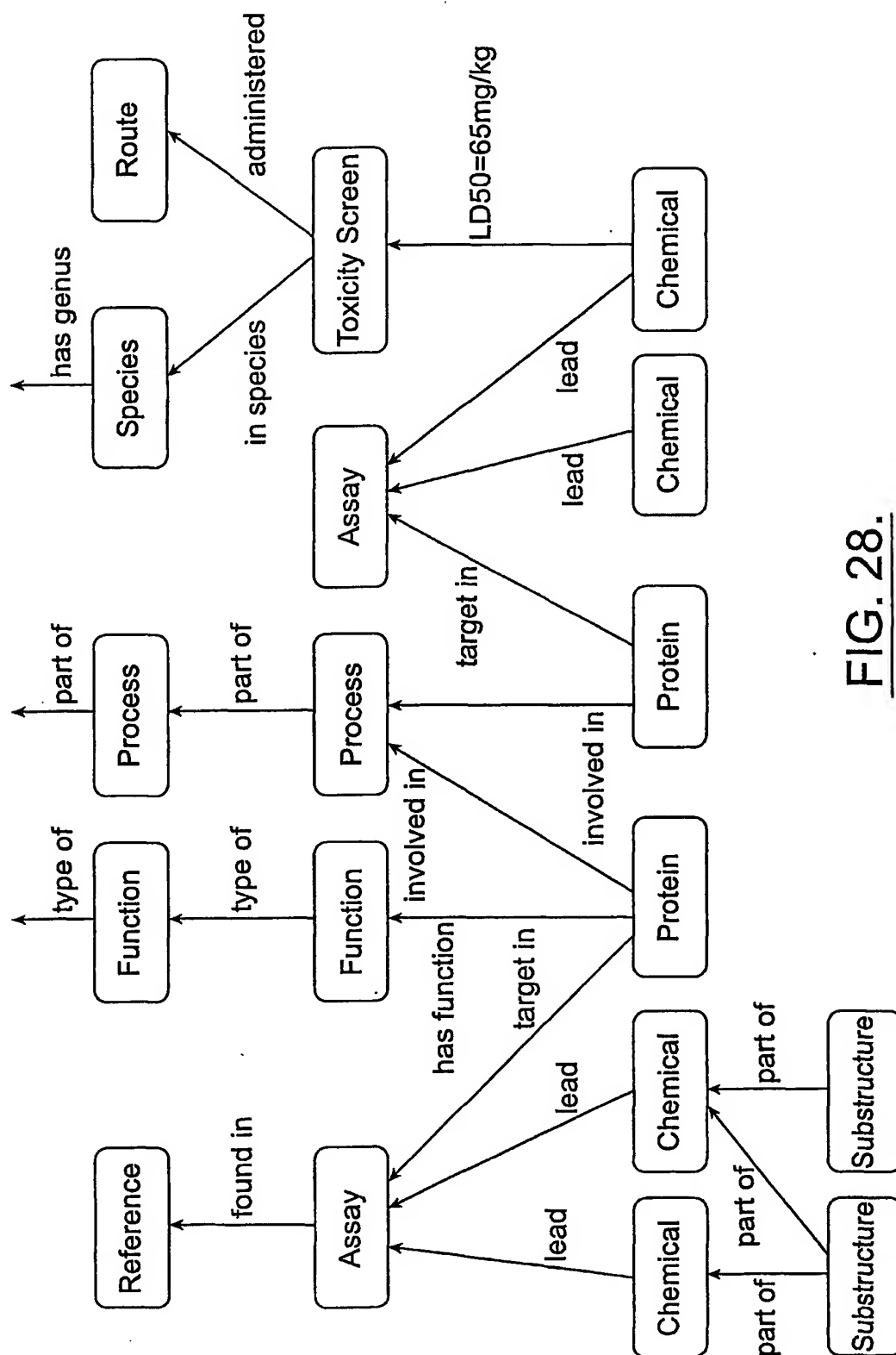


FIG. 28.

25/38

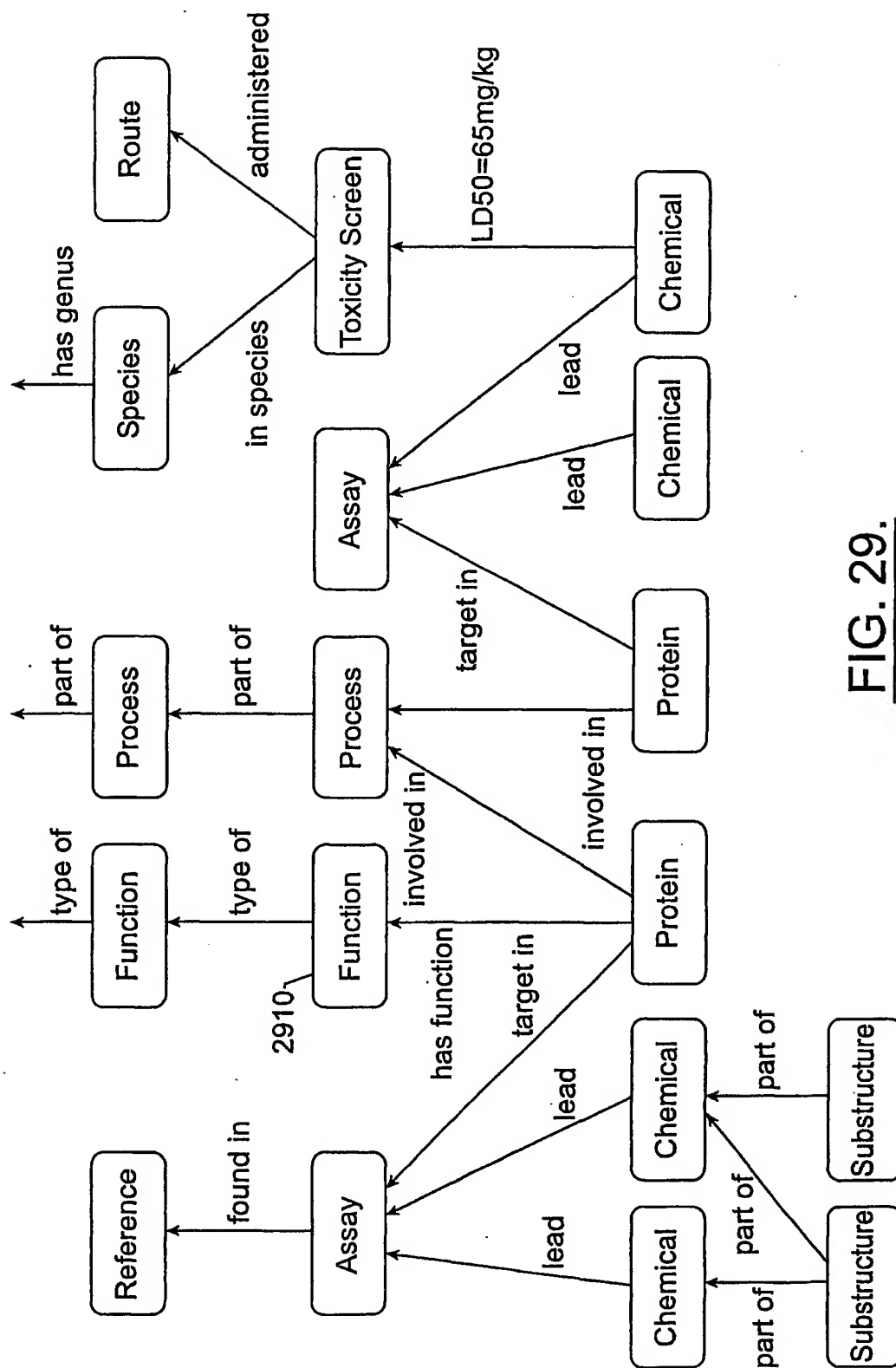


FIG. 29.

26/38

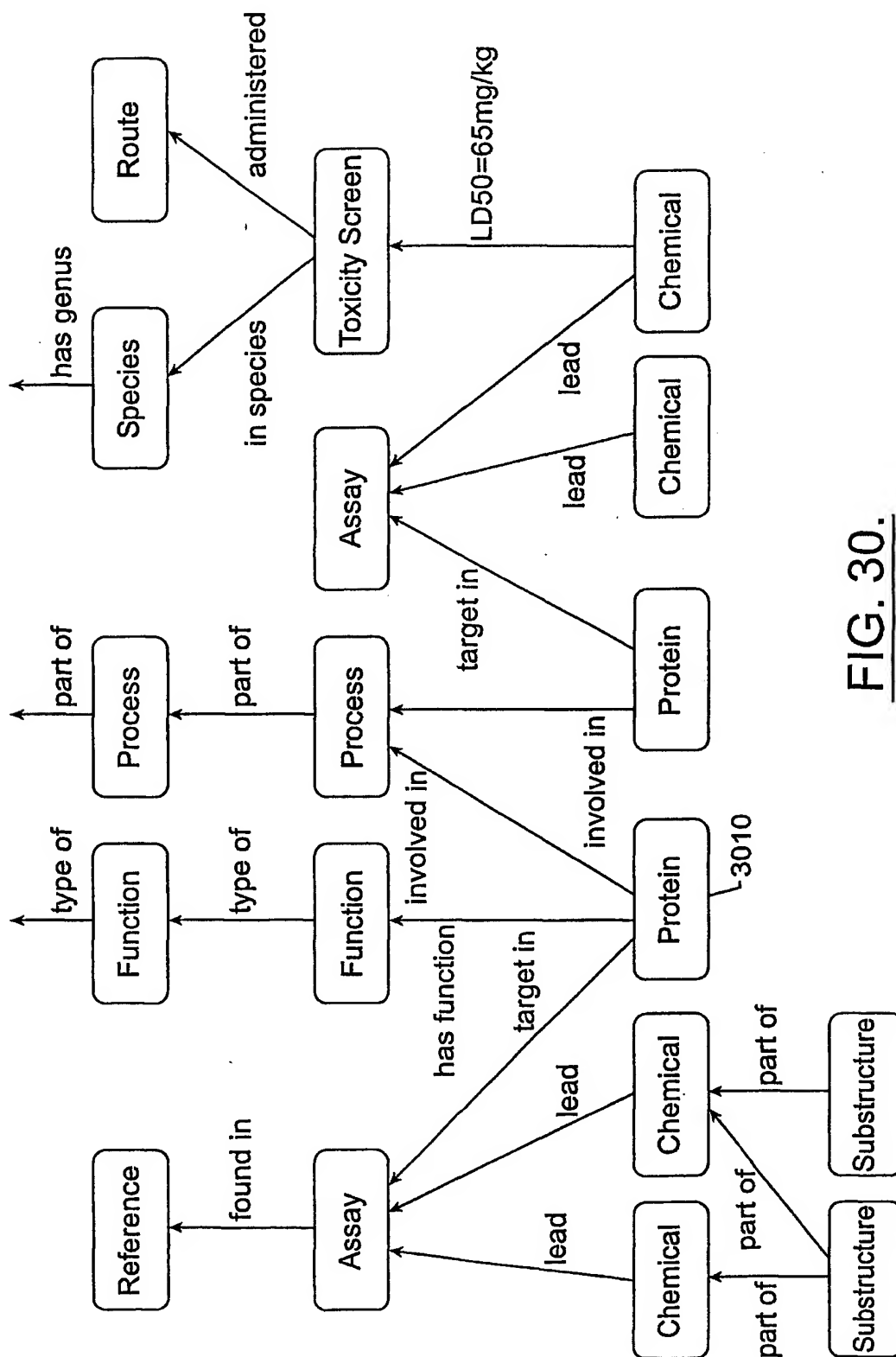


FIG. 30.

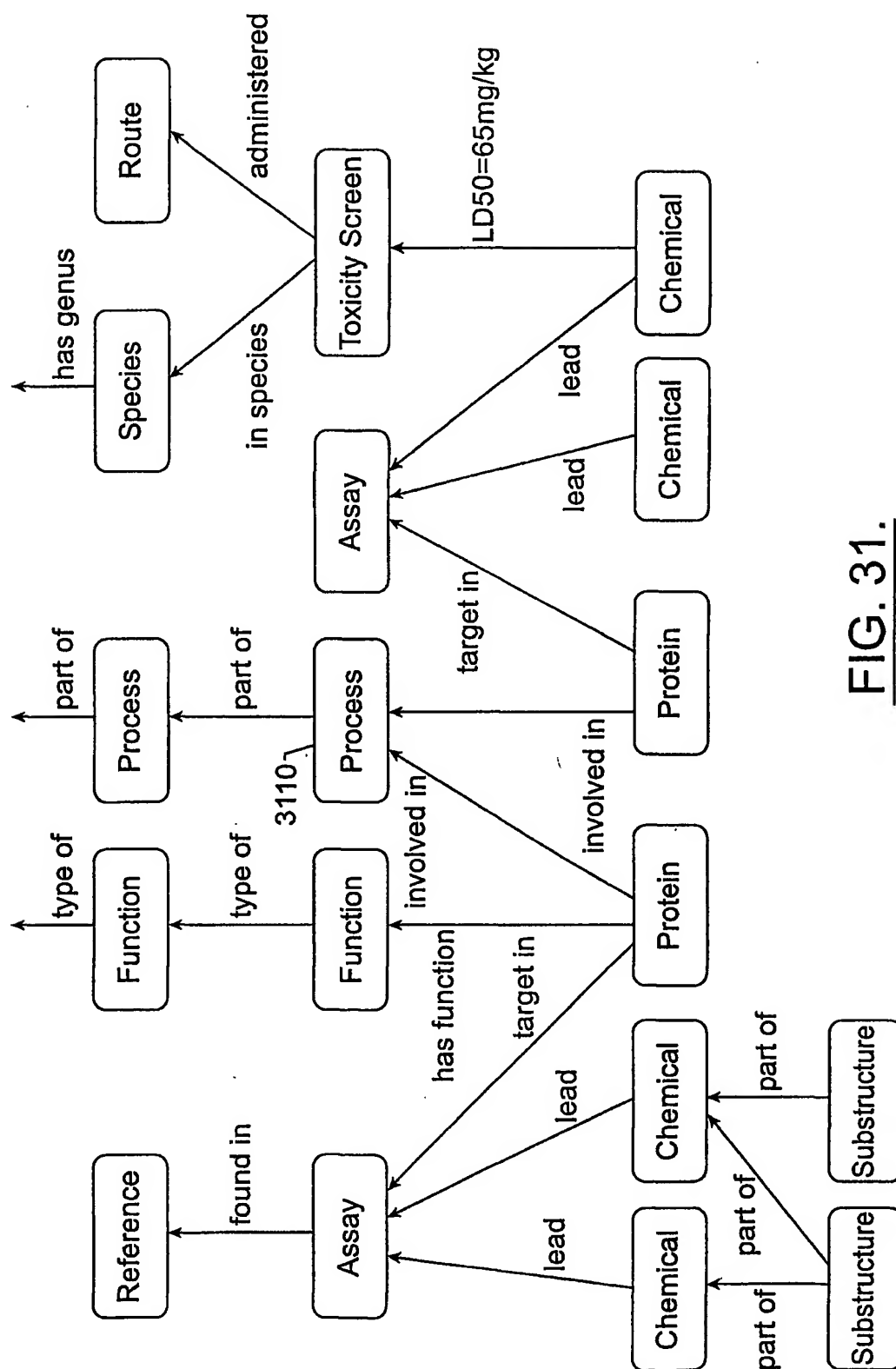


FIG. 31.

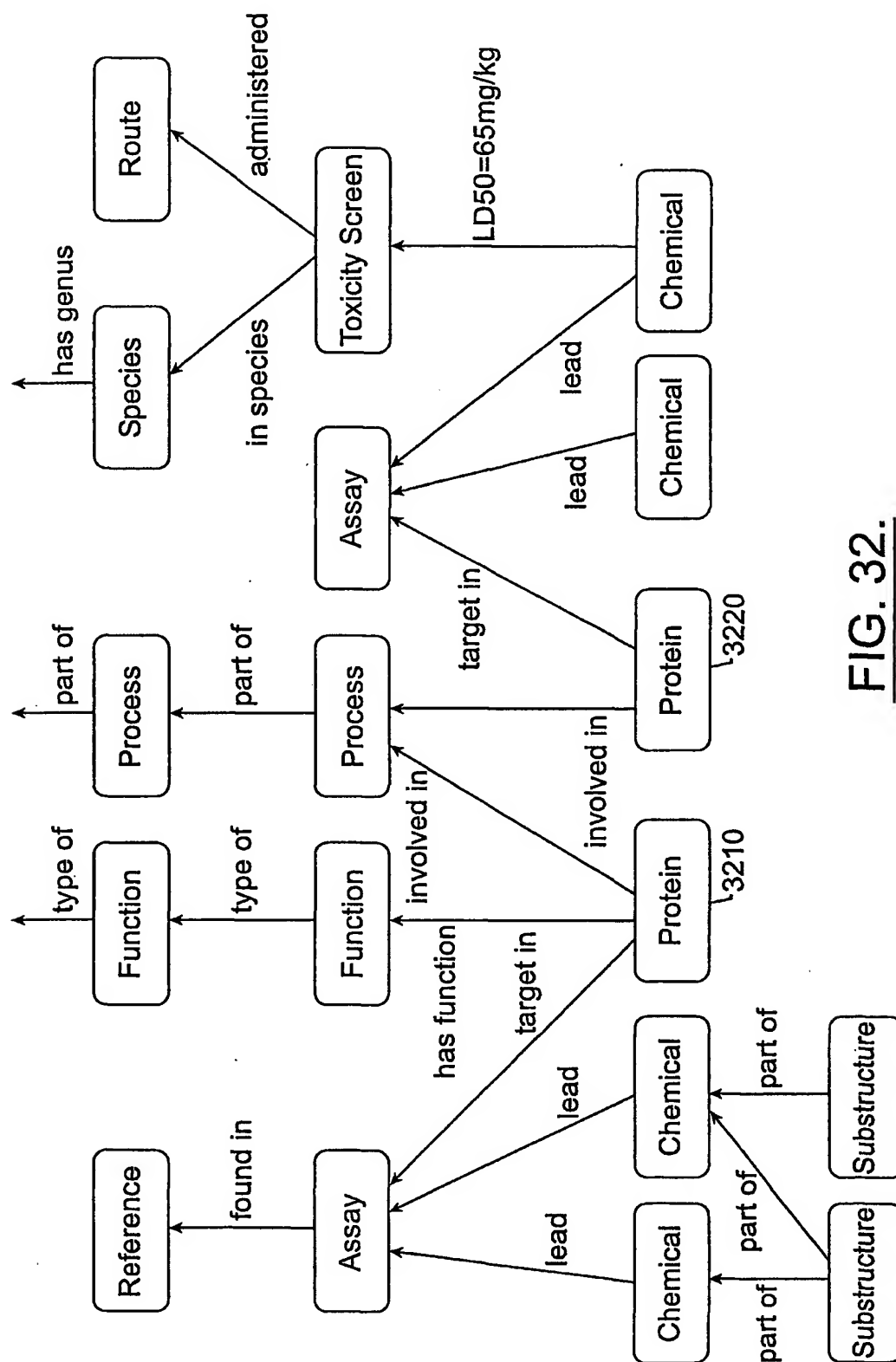


FIG. 32.

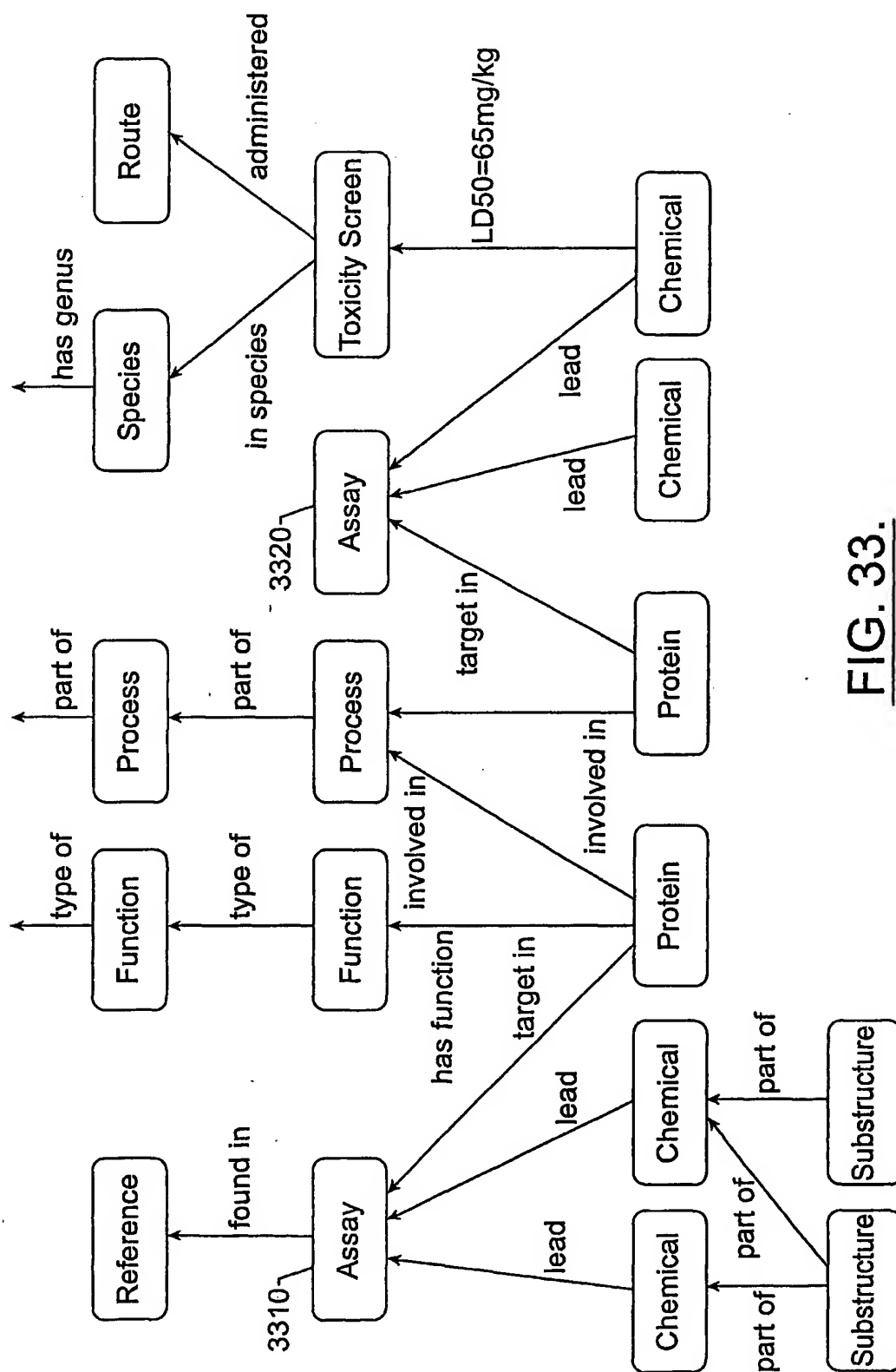


FIG. 33.

30/38

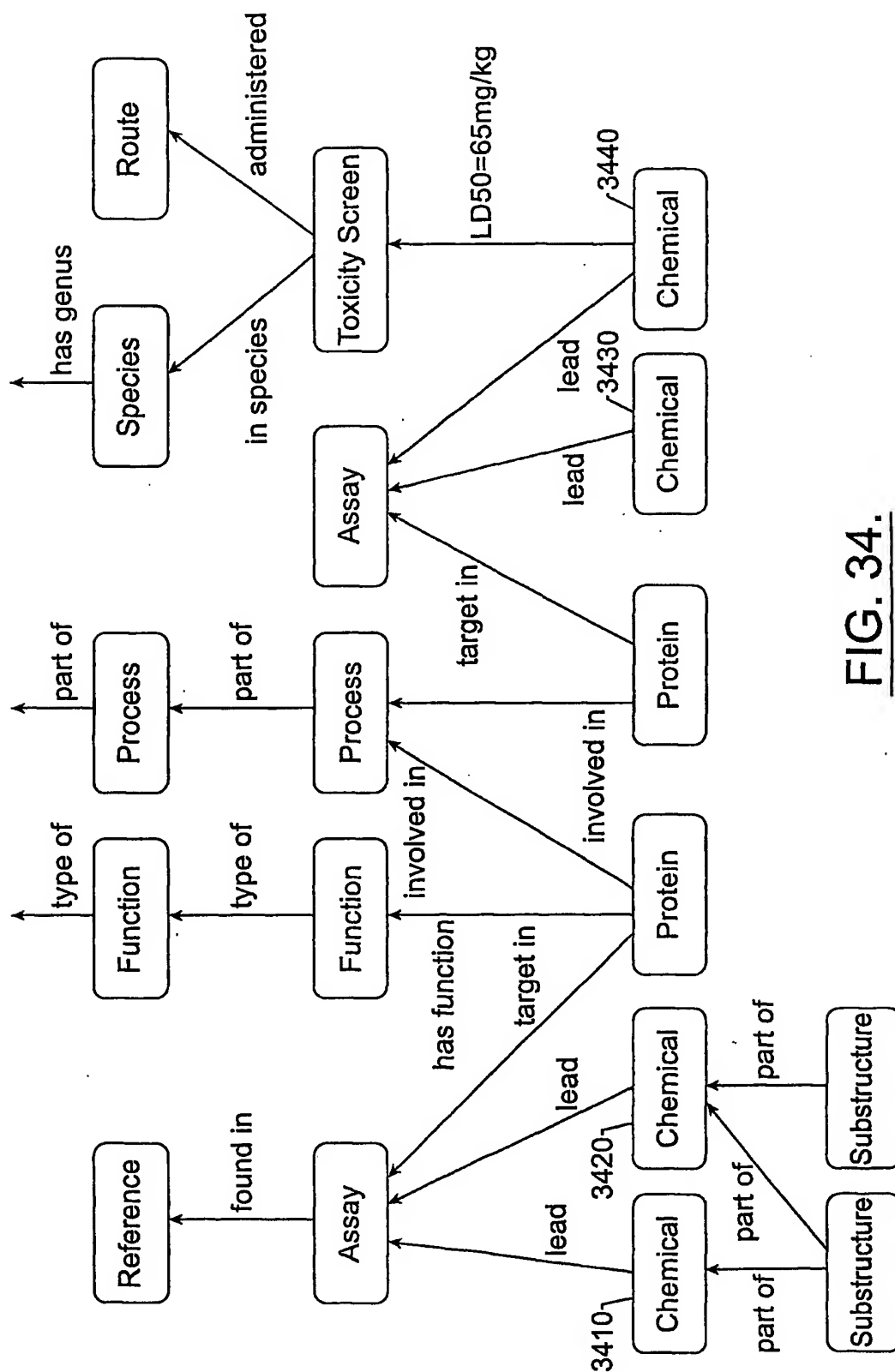


FIG. 34.

31/38

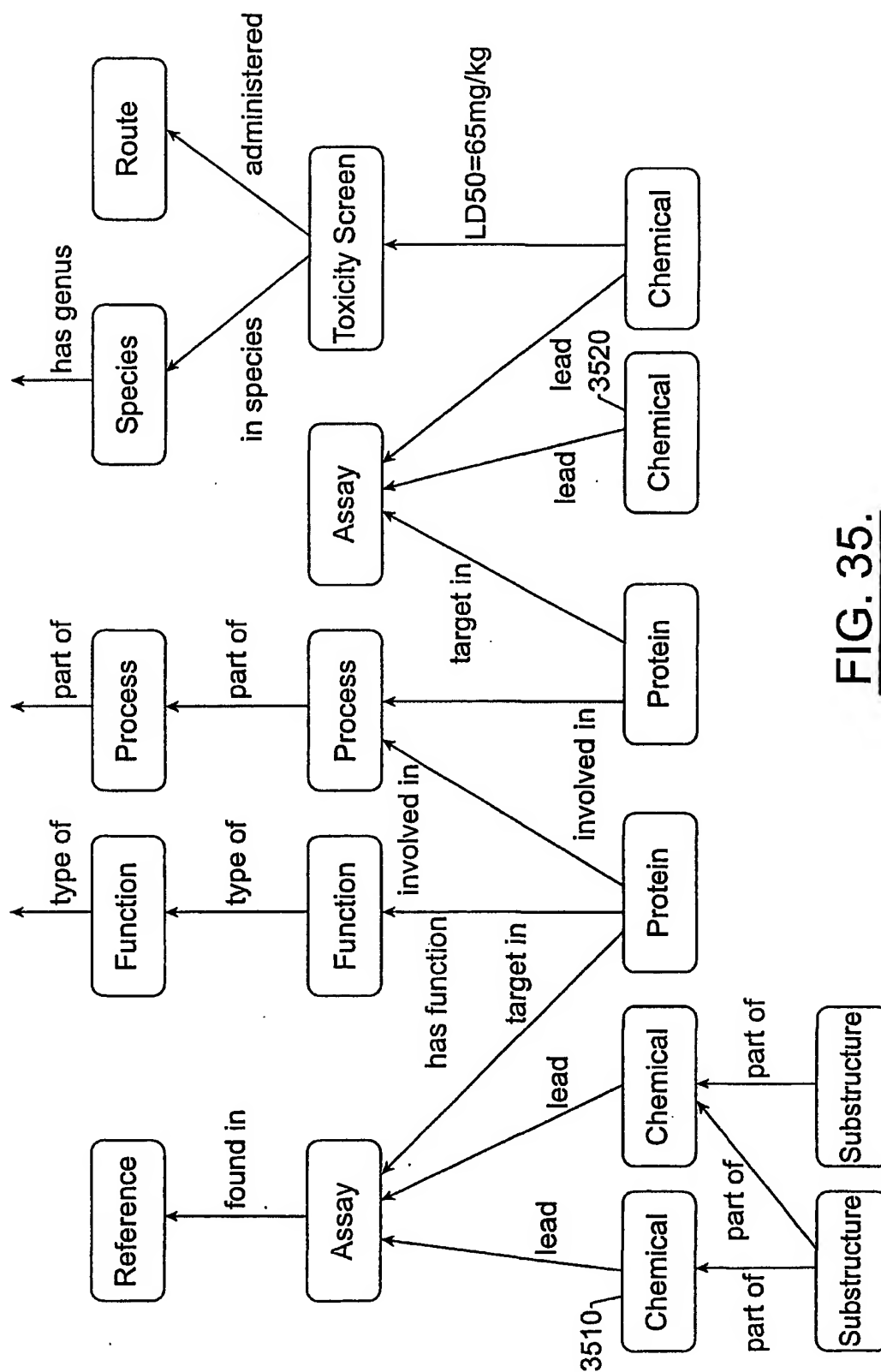


FIG. 35.

32/38

**Incellico**

Analytic software for the life sciences

Demonstration of Integrated Ontology Searching

Search for compound screening data

Target☐ SWISS-PROT ID:☒ Functional Class:**Query Expansion**☒ Expand query by process☐ Expand query by metabolic pathway**Absorption/Toxicity**☐ LD50 in species: , route: is more than mg/kg☐ Show records with no toxicity data.☐ Qualitative assessment of absorption is ▼☐ Show records with no absorption data.**Chemical Structure**☐ Must have structural elements:☐ Must not have structural elements:FIG. 36.

PROTEIN FARNESYLTRANSFERASE-1192365509-TABLE 1. IMIDAZOLE-CONTAINING PIPERAZINE INHIBITORS OF FT			
Name	CAS-ID	Result	
(S)-1-[2-(1H-IMDAXOL-4-YL)PROPYL]-2-(2-METHOXYETHYL)-4-(1-NAPHTHALENYLCARBONYL)PIPERAZINE	Unknown	IC50=10	
(S)-1-[2-(1H-IMDAXOL-4-YL)ACETYL]-2-(2-METHOXYETHYL)-4-(1-NAPHTHALENYLCARBONYL)PIPERAZINE, DIHYDROCHLORIDE	Unknown	IC50=34	
PROTEIN FARNESYLTRANSFERASE-873874829-TABLE1. INHIBITION CONSTANTS FOR FPP AND GGPP ANALOGUES			
Name	CAS-ID	Result	
3-VINYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=173	
3-TERTIARY-BUTANYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=31	
3-ALLYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=189	
3-ETHYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=215	
3-PHENYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=299	
3-VINYL-GERANYLGERANYL-DIPHOSPHATE	Unknown	IC50=715	
3-ALLYL-GERANYLGERANYL-DIPHOSPHATE	Unknown	IC50=453	
PROTEIN GERANYLGERANYLTRANSFERASE-873874829-TABLE1. INHIBITION CONSTANTS FOR FPP AND GGPP ANALOGUES			
Query Expansion: Process-C-TERMINAL/PROTEIN PRENYLATION			
Name	CAS-ID	Result	
3-VINYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=100000	
3-TERTIARY-BUTANYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=50000	
3-ALLYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=100000	
3-ETHYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=100000	
3-PHENYL-FARNESYL-DIPHOSPHATE	Unknown	IC50=100000	
3-VINYL-GERANYLGERANYL-DIPHOSPHATE	Unknown	IC50=3050	
3-ALLYL-GERANYLGERANYL-DIPHOSPHATE	Unknown	IC50=3380	

To Fig. 37B.

FIG. 37A.

From Fig. 37A.

PROTEIN GERANYLGERANYLTRANSFERASE-83554045-TABLE 1. IC50 VALUES FOR ENZYME INHIBITION, RAPA OR H-RAS PROCESSING, AND
RELATIVE SELECTIVITY FOR PGGLTASE AND PPTASEINHIBITORS 1-20
Query Expansion: Process-C-TERMINAL PROTEIN PRENYLATION

Name	CAS-ID	Result
4-NITRO-2-(1-NAPHTHYL)BENZOATE	Unknown	IC50=44
4-[N-(1H-IMIDAZOL-2-YL)METHYLENEAMINO]-2-(1-NAPHTHYL)BENZOYL]LEUCINE.	Unknown	IC50=210
4-[N-(4-PYRIDYL)METHYLENEAMINO]-2-(1-NAPHTHYL)BENZOYL]LEUCINE.	Unknown	IC50=1600
4-METHOXY-2-PHENYL BENZOATE.	Unknown	IC50=32
4-[N-(IMIDAZOL-4-YL)OMETHYLENEAMINO]-2-(1-NAPHTHYL)BENZOYL]LEUCINE.	Unknown	IC50=38
4-[N-(1-TRITYLIMIDAZOL-4-YL)METHYLENEAMINO]-2-PHENYLBENZOYL]LEUCINE METHYL ESTER.	Unknown	IC50=90

FIG. 37B.

35/38

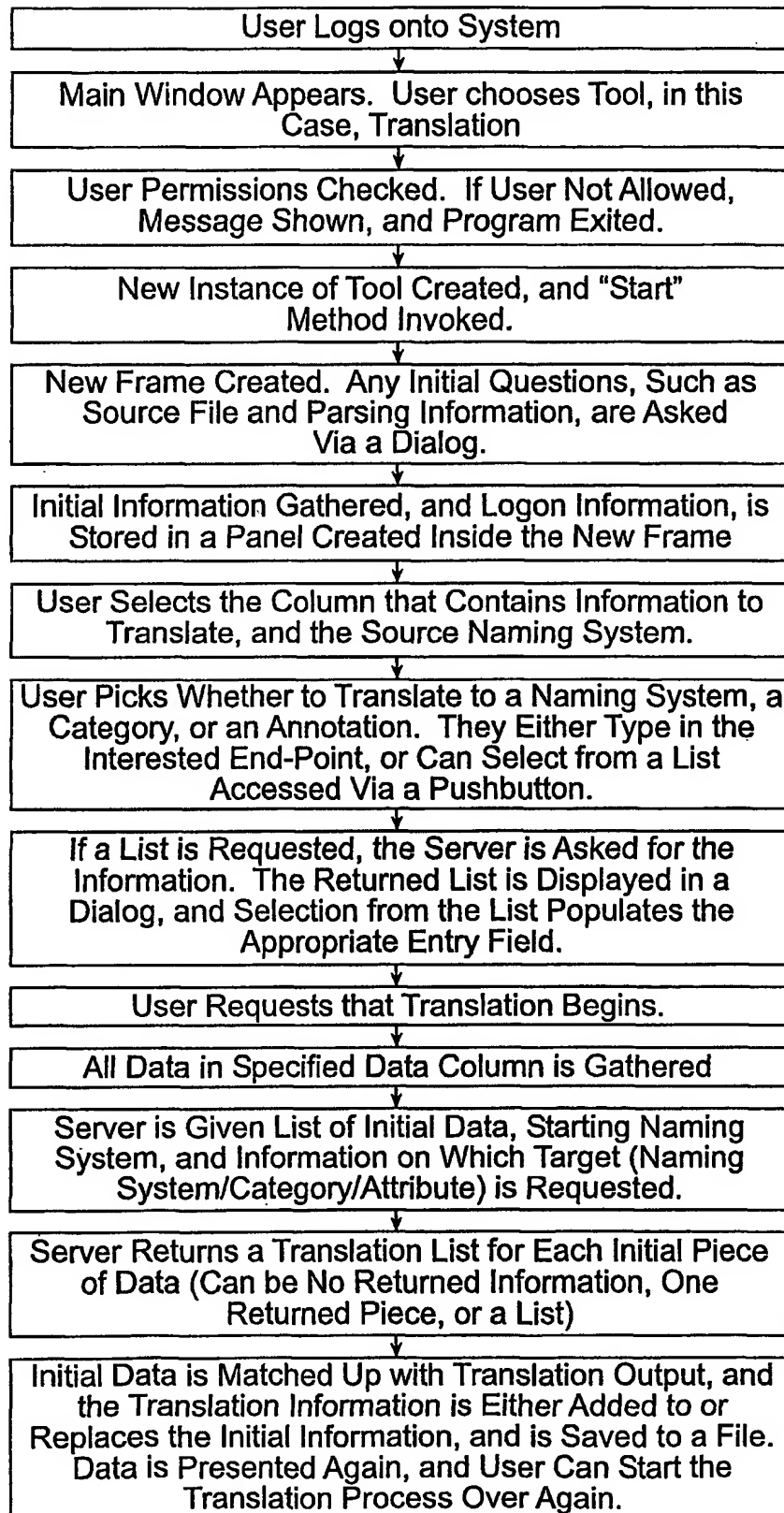
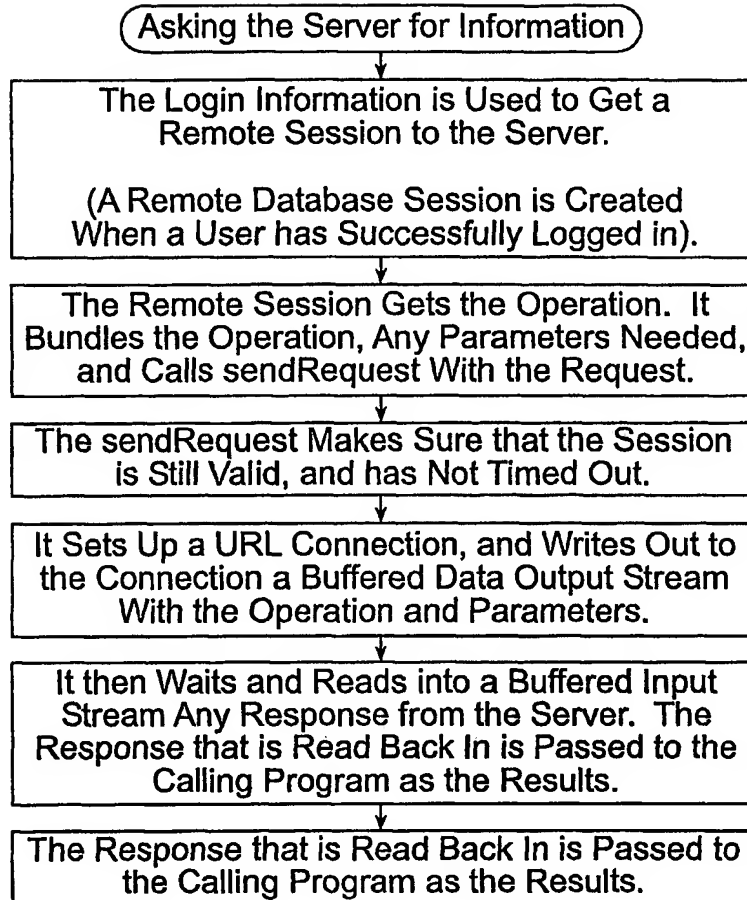
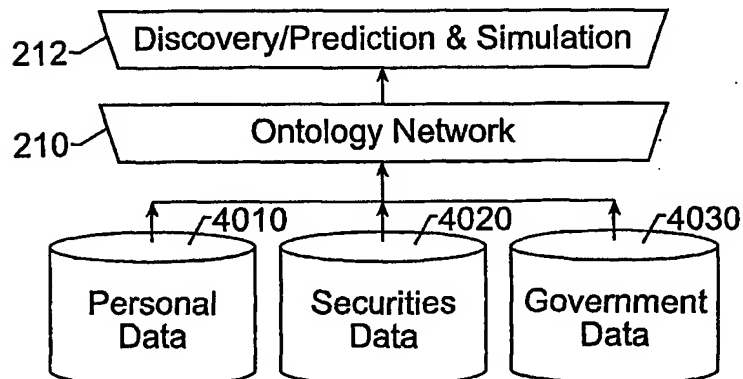


FIG. 38.

36/38

FIG. 39.FIG. 40.

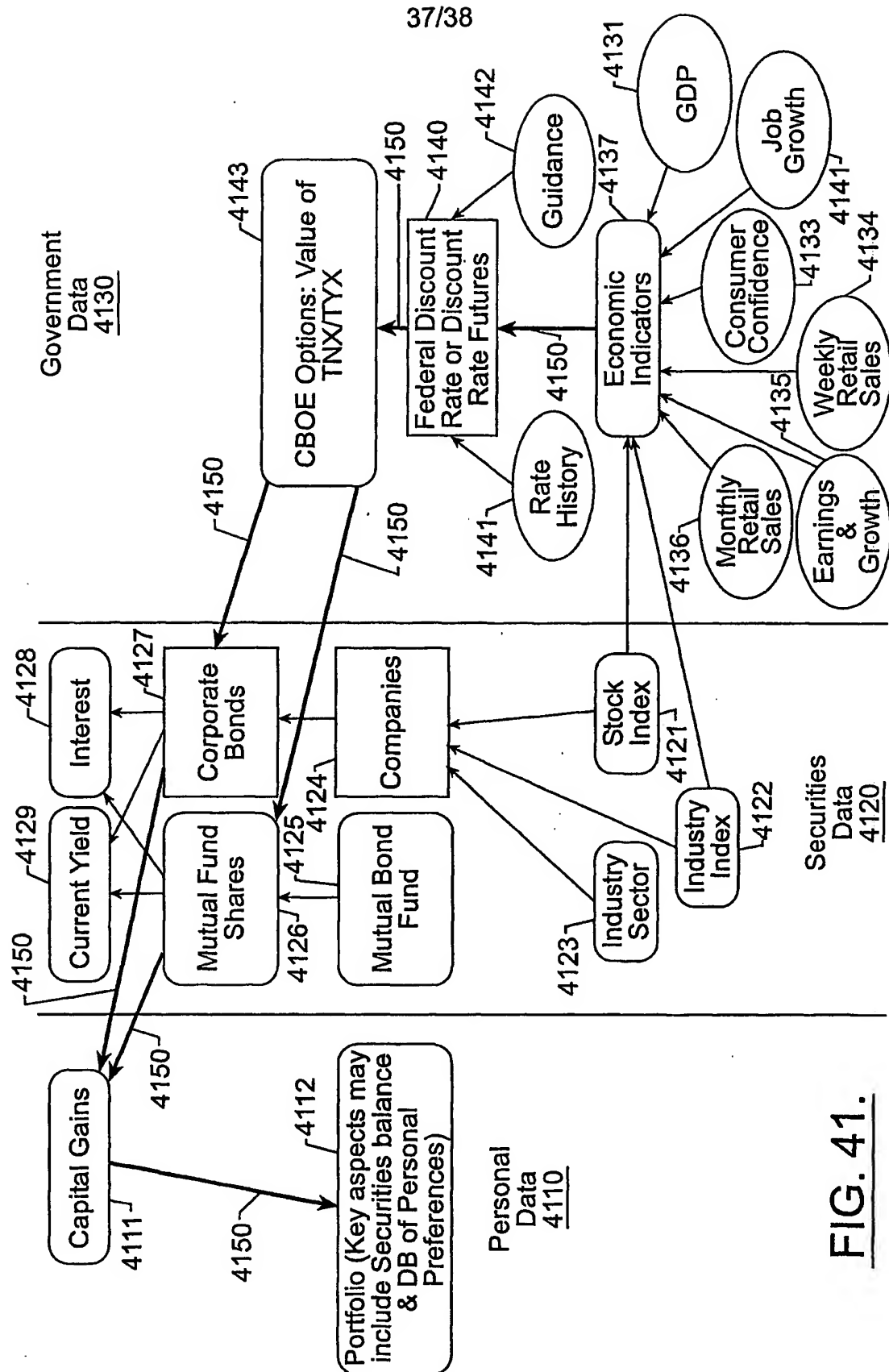


FIG. 41.

38/38

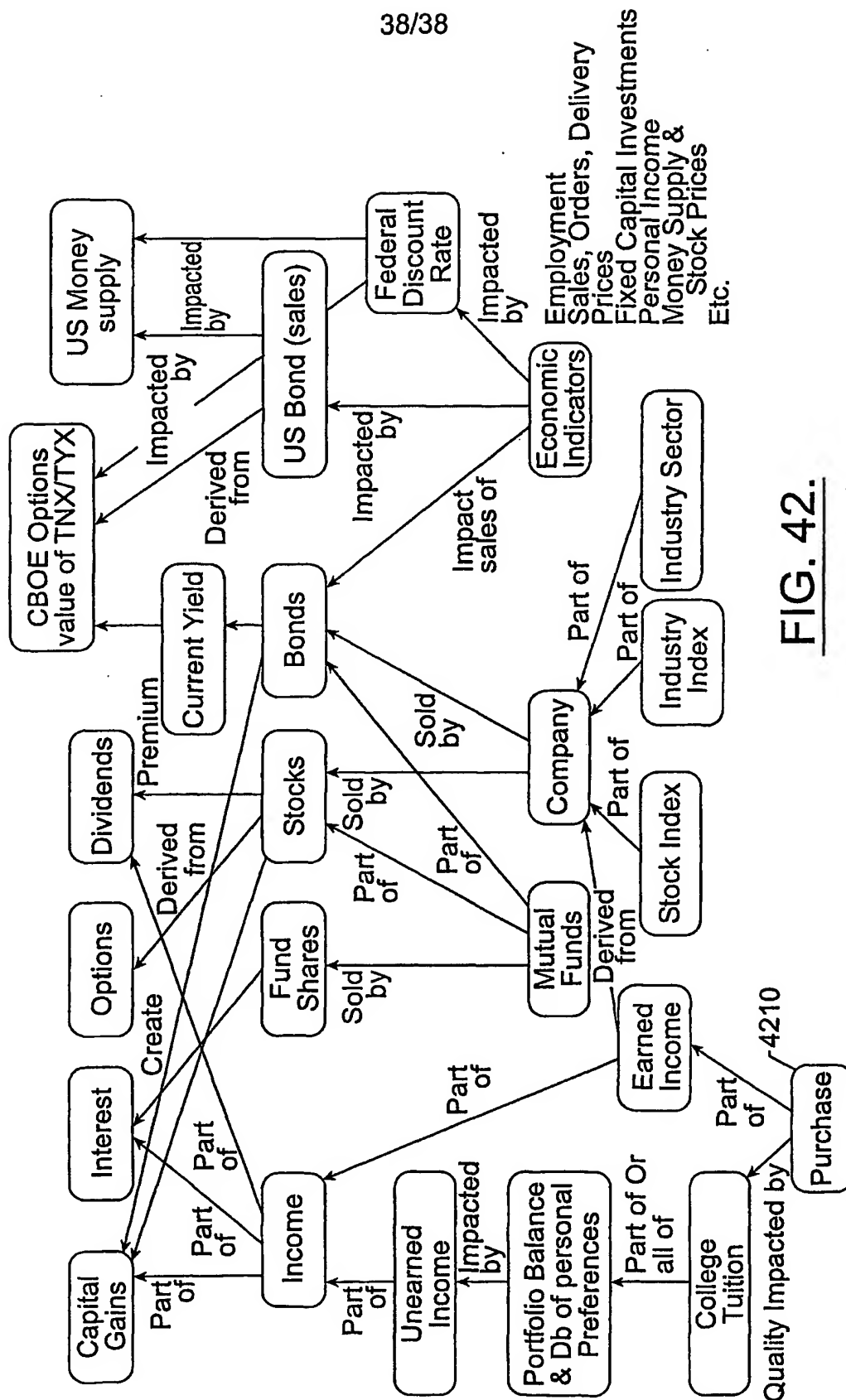


FIG. 42.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/16406

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 19/00

US CL : 702/19

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 702/19

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	XU et al. Some Experience Using OODB in the Interoperation of Heterogeneous Genome Database Systems. Proceedings of the 1996 Engineering Systems Design and Analysis Conference. July 1996, Vol. 2, pages 53-61, especially page 56.	1-115
Y	US 5,859,972 A (SUBRAMANIAM ET AL) 12 January 1999 (12.01.1999), columns 1-3.	1-115

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"B" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T"

later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X"

document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y"

document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"

document member of the same patent family

Date of the actual completion of the international search

13 July 2002 (13.07.2002)

Date of mailing of the international search report

27 AUG 2002

Name and mailing address of the ISA/US

Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703)305-3230

Authorized officer

John S. Brusca

Telephone No. 703 308-0196

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/16406

Continuation of B. FIELDS SEARCHED Item 3:

U.S. Patent, Derwent World Patent Index

search terms: search, database, link, integrate, biological, chemical

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/16406

Box I Observations where certain claims were found unsearchable (Continuation of Item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claim Nos.: 116 and 117
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 116 and 117 were not searched because they are drawn to mere representation of information which is excluded from searching under PCT Rule 39.
2. ☐ Claim Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claim Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of Item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐
☐

The additional search fees were accompanied by the applicant's protest.

No protest accompanied the payment of additional search fees.